INDUSTRIAL AND SYSTEMS ENGINEERING

ISE

A Stochastic Proximal-Gradient Method for Regularized Generally Constrained Continuous Optimization

Frank E. Curtis, Xiaoyi Qu, and Daniel P. Robinson

Department of Industrial and Systems Engineering, Lehigh University

COR@L Technical Report 26T-019

LEHIGH UNIVERSITY

COR@L COMPUTATIONAL OPTIMIZATION RESEARCH AT LEHIGH

A Stochastic Proximal-Gradient Method for Regularized Generally Constrained Continuous Optimization

Frank E. Curtis*[1], Xiaoyi Qu†[1], and Daniel P. Robinson‡[1]

[1]Department of Industrial and Systems Engineering, Lehigh University

October 26, 2026

Abstract

A stochastic proximal-gradient method is proposed and analyzed for solving regularized optimization problems subject to nonlinear equality constraints and nonnegativity bounds. The objective function's values and gradients are available only through inexact stochastic oracles. The method decomposes each trial step into one component that reduces linearized infeasibility and a second component that seeks optimality while preserving that reduction. Trial steps are accepted or rejected through a relaxed sufficient decrease condition subject to stochastic merit function evaluations. Without assuming unbiasedness of stochastic gradient estimates, it is proved under reasonable assumptions that the method reaches an ϵ-KKT point within $\mathcal{O}(\epsilon^{-2})$ iterations with high probability. This guarantee further implies almost-sure termination in finitely many iterations and an $\mathcal{O}(\epsilon^{-2})$ bound on the expected iteration complexity.

1 Introduction

In this paper, we consider the constrained continuous optimization problem

$$\min_{x\in\mathbb{R}^n} \; f(x) + r(x) \;\text{ subject to }\; c(x) = 0, \quad x \in \Omega := \{x \in \mathbb{R}^n : x \geq 0\}, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) := \mathbb{E}_{\Xi\sim\mathcal{P}}[f(x,\Xi)]$ for all $x \in \mathbb{R}^n$ is continuously differentiable, the regularization function $r : \mathbb{R}^n \to \mathbb{R}$ is convex, and the equality constraint function $c : \mathbb{R}^n \to \mathbb{R}^m$ has $m \leq n$ and is continuously differentiable. Here, the random variable Ξ is defined by a probability space $(\mathcal{E}, \mathcal{F}, \mathcal{P})$, where $\mathcal{E}$ is a sample space, $\mathcal{F}$ is a σ-algebra of measurable subsets of the sample space, and $\mathcal{P} : \mathcal{F} \to [0, 1]$ is a probability measure. For each $x \in \mathbb{R}^n$, the function $f(x, \cdot) : \mathcal{E} \to \mathbb{R}$ is measurable with respect to the σ-algebra $\mathcal{F}$ on $\mathcal{E}$. We presume that the expectation in the definition of f is approximated using samples drawn from Ξ; i.e., we assume the objective f and its derivatives are not known and must be estimated through sampling. Problems of the form (1) arise naturally in a wide range of applications, including stochastic programming [3, 22], model compression [6], and constrained deep learning [24, 25], among others. For a comprehensive list of applications, we refer readers to [12, 15, 23] and the references therein.

Recently, several algorithms have been proposed for solving problem (1), including augmented Lagrangian methods [15, 23], a proximal-point method [4], a proximal-gradient method [5], and a successive convex approximation method [12]. Their theoretical guarantees have been established in terms of expected iteration complexity, i.e., the number of iterations required for a stationarity measure to fall below a prescribed accuracy (in expectation). Results in expectation, however, do not control tail events: the iterations required to

*E-mail: frank.e.curtis@lehigh.edu
†E-mail: xiq322@lehigh.edu
‡E-mail: daniel.p.robinson@lehigh.edu

reach an approximate stationary point can still be arbitrarily large with nonzero probability. To address this limitation, a separate line of work [2, 8, 10, 13, 16, 21] has developed high-probability iteration complexity bounds that characterize the number of iterations needed to reach an approximate stationary point with a prescribed failure probability. These guarantees have been established in a variety of settings, including nonconvex stochastic optimization [8, 13, 16], nonsmooth stochastic optimization [10], equality constrained stochastic optimization [2], and stochastic variational inequality problems [21]. To the best of our knowledge, however, no previously existing method for problem (1) provides high-probability iteration complexity guarantees.

Another limitation of existing high-probability analyses [10, 13, 17] is the reliance on a stochastic gradient unbiasedness assumption. This excludes without-replacement schemes such as random reshuffling [11], which are standard in practice but yield biased mini-batch gradients within an epoch [18, 19]. Together, these limitations motivate the development of an algorithm for solving (1) that achieves high-probability iteration complexity guarantees without having to assume that the stochastic gradient estimator is unbiased.

In this paper, we design and analyze a proximal-gradient method for solving problem (1). Our method can be seen as an extension of the proximal-gradient methods in [7, 9] to the setting where the objective function and its gradient are inaccessible and must be estimated, or as an extension of the step-search method [14] to stochastic optimization problems with both nonsmooth regularization and general constraints. The proposed method employs a step decomposition strategy and uses a sufficient decrease condition for a chosen merit function to determine step acceptance or rejection. The algorithm features two key aspects.

- We propose a new stochastic first-order oracle (see Assumption 3.2) in which the required gradient-estimation accuracy is governed by both a stationarity measure and the constraint violation.
- To address the challenges arising from noise in objective function evaluations, we employ a relaxed sufficient decrease condition similar to [14]. Under reasonable assumptions, we prove that, with high probability, the proposed algorithm generates an ϵ-KKT point within $\mathcal{O}(\epsilon^{-2})$ iterations. Furthermore, the high-probability result implies both an almost-sure termination guarantee (Corollary 3.19(i)) and an expected iteration complexity bound (Corollary 3.19(ii)).

Before proceeding, it is prudent that we clarify the contributions of this paper relative to two other papers in the literature.

- The problem studied in [7] is similar to our problem (1) but they assume that the function f and its gradient are known exactly. To address inexactness, we use a modified line search and an entirely different analysis (see next bullet).
- The problem studied in [14] is an *unconstrained* optimization problem where the objective function is assumed to be differentiable, and possibly nonconvex. Our problem (1) is distinct since it allows for the use of a nonsmooth regularization function and the presence of constraints. The overall framework for analyzing their method is the same framework that we adopt. However, since the two papers consider very different optimization problems, and the optimization algorithms analyzed are completely different, we overcome various challenges in our analysis that were not an issue in [14], such as needing to define an appropriate first-order oracle that takes constraint infeasibility into account and defining and analyzing an appropriate stopping time that considers both the constraints and nonsmooth regularizer.

1.1 Organization

In Section 2, we provide the description of the proximal-gradient algorithm. In Section 3, we derive a high-probability iteration complexity result, which implies both an almost-sure termination result and expected iteration complexity result. Finally, we give concluding remarks in Section 4.

1.2 Notation

Let $\mathbb{R}$ denote the set of real numbers, $\mathbb{R}^n$ the set of n-dimensional real vectors, and $\mathbb{N}$ the set of natural numbers. For any $a \in \mathbb{R}$, let $\mathbb{R}_{>a}$ ($\mathbb{R}_{\geq a}$) denote the set of real numbers strictly larger than (larger than or

equal to) a. We use $k \in \mathbb{N}$ as the iteration counter of the algorithm. Given a matrix A, we let $\sigma_{\min}(A)$ denote the smallest singular value of A. Finally, we let $\exp(a) := e^a$.

1.3 Standing assumption

The following assumption is made throughout the paper.

Assumption 1.1. *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be an open convex set containing every realization of the iterate, iterate plus normal step, and iterate plus full step (i.e., trial iterate) sequences of the algorithm. The objective function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and bounded below over $\mathcal{X}$. The objective gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous over $\mathcal{X}$. The regularization function $r : \mathbb{R}^n \to \mathbb{R}$ is convex, Lipschitz continuous, and bounded below over $\mathcal{X}$. The constraint function $c : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable and bounded over $\mathcal{X}$, and constraint gradient $\nabla c_i : \mathbb{R}^n \to \mathbb{R}^n$ for all $i \in \{1, \ldots, m\}$ is Lipschitz continuous and bounded over $\mathcal{X}$.*

Under Assumption 1.1, there exist constants $(f_{\inf}, r_{\inf}, \kappa_{\partial r}, \kappa_c, \kappa_J, L_g, L_J) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ such that for all $x \in \mathcal{X}$, one has

$$f(x) \geq f_{\inf},\ r(x) \geq r_{\inf},\ d_H(\partial r(x), 0) \leq \kappa_{\partial r},\ \|c(x)\|_2 \leq \kappa_c,\ \text{and}\ \|\nabla c(x)\|_2 \leq \kappa_J \tag{2}$$

where $d_H(\partial r(x), 0)$ denotes the ℓ_2-norm Hausdorff distance between the set $\partial r(x)$ and the origin set $\{0\}$, and for all $(x, \bar{x}) \in \mathcal{X} \times \mathcal{X}$, one has

$$\|\nabla f(x) - \nabla f(\bar{x})\|_2 \leq L_g \|x - \bar{x}\|_2 \ \text{ and } \ \|\nabla c(x)^T - \nabla c(\bar{x})^T\|_2 \leq L_J \|x - \bar{x}\|_2. \tag{3}$$

All parameters used in this paper (including those above) are gathered in Table 1.

Table 1: Parameters and where they are defined.

Param.	Def.	Param.	Def.	Param.	Def.
$\mathcal{X}$	Ass. 1.1	$f_{\inf}$, $r_{\inf}$, $\kappa_{\partial r}$, κ_c, κ_J	(2)	L_g, L_J	(3)
γ, $\bar{\epsilon}_f$	Alg. 1	κ_G	Ass. 3.5	ω_1, ω_2, κ_T, $\bar{\alpha}$	Def. 3.5
κ_v	(4)	$\sigma_{\min}$	Ass. 3.4	κ_1	Lem. 3.1
η, $\bar{i}$, θ	Def. 3.5	ν, b, ϵ_f	Ass. 3.1	ϵ_g, κ, δ	Ass. 3.2
u_f, p	Ass. 3.3	η_1	(9)	$\overline{W}_0$, $\hat{t}$	Thm. 3.18
ϵ_τ	(10)	$\tau_{\min}$	Lem. 3.3		

2 Algorithm

Our proposed method is stated as Algorithm 1 on page 6. For ease of development, we state the algorithm in the context of a realized iterate sequence $\{x_k\}$ and the computations that are performed for each k. In the following section, we analyze the algorithm with respect to the stochastic process that it defines.

During the kth iteration, a trial step $s_k = v_k + u_k$ is computed, where v_k reduces linearized infeasibility of the equality constraints while maintaining feasibility with respect to Ω, and u_k aims to reduce the objective function while preserving the gains achieved by v_k. We then decide whether to accept the new trial iterate $x_k + s_k$ by considering the decrease in a noisy estimate of a merit function.

Given the kth iterate x_k and proximal parameter α_k, the normal step v_k is computed as an approximate solution to the trust-region subproblem

$$\min_{v \in \mathbb{R}^n} m_k(v) \ \text{ subject to } \ \|v\|_2 \leq \kappa_v \alpha_k \delta_k, \quad x_k + v \in \Omega, \tag{4}$$

where $m_k(v) := \frac{1}{2}\|c(x_k) + \nabla c(x_k)^T v\|_2^2$, $\kappa_v \in \mathbb{R}_{>0}$ is a user-defined constant, and

$$\delta_k := \|\mathrm{Proj}_{T_\Omega(x_k)}(-\nabla c(x_k)c(x_k))\|_2. \tag{5}$$

Here, $T_\Omega(x_k)$ is the tangent cone to Ω at x_k (see [7, Definition 2.4] for details) and $\mathrm{Proj}_{T_\Omega(x_k)}(\cdot)$ is the orthogonal projection operator onto the set $T_\Omega(x_k)$. An approximate solution v_k to subproblem (4) is required to satisfy

$$\|v_k\|_2 \le \kappa_v \alpha_k \delta_k, \quad x_k + v_k \in \Omega, \quad \text{and} \quad m_k(v_k) \le m_k(v_k^c), \tag{6}$$

where v_k^c denotes the Cauchy point of subproblem (4) (see [7, Section 3]). Throughout the paper, the computation of v_k is assumed to satisfy the following.

Assumption 2.1. *Given an iterate x_k and proximal parameter α_k, a deterministic procedure is used to compute a normal step v_k satisfying* (6)*.*

Given the kth iterate x_k, proximal parameter α_k, normal step v_k, and gradient estimate $g_k \approx \nabla f(x_k)$, the step u_k is computed as the unique solution to

$$\min_{u \in \mathbb{R}^n} \quad q_k(v_k + u) \;\; \text{subject to} \;\; \nabla c(x_k)^T u = 0, \quad x_k + v_k + u \in \Omega, \tag{7}$$

with $q_k(v_k + u) := g_k^T(v_k + u) + \frac{1}{2\alpha_k}\|v_k + u\|_2^2 + r(x_k + v_k + u)$, where the constraint $\nabla c(x_k)^T u = 0$ ensures that u_k lies in the null space of $\nabla c(x_k)^T$, so any gain in linearized feasibility achieved by v_k is preserved. From the optimality conditions for (7), there exist $g_{r,k} \in \partial r(x_k + v_k + u_k)$, $y_k \in \mathbb{R}^m$, and $z_k \in \mathbb{R}^n$ such that

$$g_k + \tfrac{1}{\alpha_k} u_k + \tfrac{1}{\alpha_k} v_k + g_{r,k} + \nabla c(x_k) y_k + z_k = 0, \tag{8a}$$

$$\nabla c(x_k)^T u_k = 0, \quad \text{and} \tag{8b}$$

$$\|\min\{x_k + v_k + u_k, -z_k\}\|_2 = 0. \tag{8c}$$

We write $s_k = v_k + u_k$ for the trial step. The next assumption complements Assumption 2.1 and concerns the computation of $g_{r,k}$, y_k, and z_k.

Assumption 2.2. *Given an iterate x_k, proximal parameter α_k, and normal step v_k satisfying Assumption 2.1, a deterministic procedure is used to compute the vectors u_k, $g_{r,k} \in \partial r(x_k + v_k + u_k)$, y_k, and z_k satisfying* (8)*.*

Upon computation of the aforementioned quantities, an update is considered in the value of a merit parameter. Toward this end, the algorithm first sets, for a user-defined parameter $\eta_1 \in (0, 1/3)$, the quantity

$$\tau_{k,\text{trial}} \leftarrow \begin{cases} \infty & \text{if } q_k(s_k) - q_k(0) \le 0, \\ \frac{(1-2\eta_1)(\|c(x_k)\|_2 - \|c(x_k) + \nabla c(x_k)^T s_k\|_2)}{q_k(s_k) - q_k(0)} & \text{otherwise,} \end{cases} \tag{9}$$

and then computes, for some $\epsilon_\tau \in (0, 1)$, the value of the kth merit parameter as

$$\tau_k \leftarrow \begin{cases} \tau_{k-1} & \text{if } \tau_{k-1} \le \tau_{k,\text{trial}}, \\ \min\{(1 - \epsilon_\tau)\tau_{k-1}, \tau_{k,\text{trial}}\} & \text{otherwise.} \end{cases} \tag{10}$$

Subsequently, the algorithm estimates whether the trial iterate $x_k + s_k$ decreases the merit function

$$\phi(x, \tau) = \tau\big(f(x) + r(x)\big) + \|c(x)\|_2, \tag{11}$$

which balances the objective function with violation of the equality constraints. Since exact evaluation of f is assumed to be impossible, the algorithm computes independent estimates $f_k \approx f(x_k)$ and $f_k^+ \approx f(x_k + s_k)$, which in turn allows it to compute $\phi_k \approx \phi(x_k, \tau_k)$ and $\phi_k^+ \approx \phi(x_k + s_k, \tau_k)$. The trial step acceptance test in Line 11 compares the change in the stochastic merit function values against a certain predicted model decrease, with a tolerance of $2\bar{\epsilon}_f$ added to account for the noise in the estimates of f. If the test passes, the trial step is accepted and the value of the proximal parameter is increased by a factor γ^{-1} subject to imposing that it never grows larger than α_0; otherwise, the iterate remains unchanged and the value of the proximal parameter is reduced by a factor γ.

Algorithm 1 Stochastic Proximal-Gradient Method

Input: $x_0 \in \Omega$, $(\alpha_0, \tau_{-1}) \in (0,1]^2$, $\bar{\epsilon}_f \in \mathbb{R}_{>0}$, $\eta_1 \in (0, 1/3)$, and $(\gamma, \epsilon_\tau) \in (0,1)^2$.
for $k = 0, 1, 2, \ldots$ **do**
  Compute v_k satisfying (6) as an approximate solution to subproblem (4).
  Generate an estimate $g_k \approx \nabla f(x_k)$.
  Compute u_k as the unique solution to subproblem (7).
  Set $s_k \leftarrow v_k + u_k$.
  Set τ_k using (10).
  Set $x_k^+ \leftarrow x_k + s_k$.
  Compute estimates $f_k \approx f(x_k)$ and $f_k^+ \approx f(x_k^+)$.
  Set $\phi_k \leftarrow \tau_k\big[f_k + r(x_k)\big] + \|c(x_k)\|_2$ and $\phi_k^+ \leftarrow \tau_k\big[f_k^+ + r(x_k^+)\big] + \|c(x_k^+)\|_2$.
  if $\phi_k^+ - \phi_k \leq -\eta_1\Big[\tau_k \frac{\|s_k\|_2^2}{2\alpha_k} + (\|c(x_k)\|_2 - \|c(x_k) + \nabla c(x_k)^T v_k\|_2)\Big] + 2\bar{\epsilon}_f$ **then**
    Set $x_{k+1} \leftarrow x_k^+$ and $\alpha_{k+1} \leftarrow \min\{\alpha_0, \gamma^{-1}\alpha_k\}$.
  else
    Set $x_{k+1} \leftarrow x_k$ and $\alpha_{k+1} \leftarrow \gamma\alpha_k$.
  end if
end for

3 Analysis

We now commence our analysis of Algorithm 1. In Section 3.1, we describe the stochastic process defined by Algorithm 1 along with related assumptions and definitions, and then analyze a stopping time for the process in Section 3.2.

3.1 Stochastic Process and Related Assumptions and Definitions

Algorithm 1 uses estimates $g_k \approx \nabla f(x_k)$, $f_k \approx f(x_k)$, and $f_k^+ \approx f(x_k + s_k)$ during each iteration k. Therefore, to analyze the algorithm we must first describe how these estimates are computed and the assumptions we make about these estimates, and then introduce notation that describes the overall stochastic process.

We begin by defining the stochastic process primarily in terms of quantities that are actually computed by the algorithm, and later introduce additional sequences of random variables that are defined by the process that are useful for analyzing it. For each $k \in \mathbb{N}$, let $M_k = (\Xi_{k,g}, \Xi_{k,f}, \Xi_{k,f}^+)$ be a triple of independent random variables each with the same distribution as Ξ used to define $f(\cdot)$. From this triple, we presume that the algorithm makes use of samples $(\xi_{k,g}, \xi_{k,f}, \xi_{k,f}^+)$ and defines

$$g_k := \nabla f(x_k, \xi_{k,g}), \quad f_k := f(x_k, \xi_{k,f}), \quad \text{and} \quad f_k^+ := f(x_k + s_k, \xi_{k,f}^+). \tag{12}$$

We define the associated estimation errors as

$$e_k := |f_k - f(x_k)| \quad \text{and} \quad e_k^+ := |f_k^+ - f(x_k^+)|. \tag{13}$$

For the analysis only, it will be convenient to define the following quantities.

Definition 3.1 (True primal-dual solution)**.** *For each $k \in \mathbb{N}$, given x_k and v_k, the true primal-dual solution to (7), i.e., the one obtained when g_k is replaced by $\nabla f(x_k)$, is $(u_k^{\text{true}}, g_{r,k}^{true}, y_k^{\text{true}}, z_k^{\text{true}})$, and in addition $s_k^{\text{true}} := v_k + u_k^{\text{true}}$.*

Associated with $\{M_k\}_{k\geq 0}$, we have filtrations $\{\mathcal{F}_k\}_{k\geq 0}$ and $\{\mathcal{F}_{k+1/3}\}_{k\geq 0}$ where

$$\mathcal{F}_0 = \sigma(x_0, \alpha_0), \qquad \mathcal{F}_k = \sigma(x_0, \alpha_0, M_0, M_1, \ldots, M_{k-1}) \ \text{ for } k \geq 1,$$
$$\mathcal{F}_{1/3} = \sigma(x_0, \alpha_0, \Xi_{0,g}), \qquad \mathcal{F}_{k+1/3} = \sigma(x_0, \alpha_0, M_0, M_1, \ldots, M_{k-1}, \Xi_{k,g}) \ \text{ for all } k \geq 1,$$

where x_0 is the initial iterate (presumed fixed), α_0 is the initial proximal parameter (also fixed), and $\sigma(\cdot)$ is the sigma algebra generated by its arguments. Overall, we analyze the stochastic process defined, for all $k \in \mathbb{N}$, by the random variables

$$\begin{aligned}
&\{X_k, \Delta_k, V_k^c, V_k, A_k\}, \text{ which are } \mathcal{F}_k\text{-measurable,}\\
&\{G_k, U_k, U_k^{\text{true}}, S_k, S_k^{\text{true}}, Y_k, Y_k^{\text{true}}, Z_k, Z_k^{\text{true}}, G_{r,k}, G_{r,k}^{\text{true}}, \mathcal{T}_{k,\text{trial}}, \mathcal{T}_k, X_k^+\},\\
&\qquad\qquad \text{which are } \mathcal{F}_{k+1/3}\text{-measurable, and}\\
&\{\Phi_k, \Phi_k^+, F_k, F_k^+, E_k, E_k^+\}, \text{ which are } \mathcal{F}_{k+1}\text{-measurable,}
\end{aligned}$$

where for each $k \in \mathbb{N}$ these quantities have the aforementioned realizations

$$\begin{aligned}
&\{x_k, \delta_k, v_k^c, v_k, \alpha_k\},\\
&\{g_k, u_k, u_k^{\text{true}}, s_k, s_k^{\text{true}}, y_k, y_k^{\text{true}}, z_k, z_k^{\text{true}}, g_{r,k}, g_{r,k}^{\text{true}}, \tau_{k,\text{trial}}, \tau_k, x_k^+\}, \text{ and}\\
&\{\phi_k, \phi_k^+, f_k, f_k^+, e_k, e_k^+\}.
\end{aligned}$$

We make the following pair of assumptions. For all $k \in \mathbb{N}$, we write $\mathbb{P}_\bullet[\cdot|\mathcal{F}_k]$ to denote probability of event $\cdot$ with respect to the distribution of the random variable $\bullet$ conditioned on $\mathcal{F}_k$, and similarly define $\mathbb{E}_\bullet[\cdot|\mathcal{F}_k]$ as the expected value of $\cdot$ with respect to the distribution of the random variable $\bullet$ conditioned on $\mathcal{F}_k$.

Assumption 3.1 (Stochastic zeroth-order oracle)**.** *There exists $(\nu, b, \epsilon_f) \in \mathbb{R}^3_{>0}$ such that, for all $k \in \mathbb{N}$, the estimation error at X_k satisfies*

$$\begin{aligned}
&\mathbb{E}_{\Xi_{k,f}}[E_k|\mathcal{F}_k] \le \epsilon_f \quad \textit{and}\\
&\mathbb{E}_{\Xi_{k,f}}\left[\exp\left\{\lambda\big(E_k - \mathbb{E}_{\Xi_{k,f}}[E_k|\mathcal{F}_k]\big)\right\}|\mathcal{F}_k\right] \le \exp\left(\frac{\lambda^2\nu^2}{2}\right) \quad \textit{for all} \;\; \lambda \in \left[0, \frac{1}{b}\right],
\end{aligned} \tag{14}$$

and the estimation error at X_k^+ satisfies

$$\begin{aligned}
&\mathbb{E}_{(\Xi_{k,g},\Xi_{k,f}^+)}[E_k^+|\mathcal{F}_k] \le \epsilon_f \quad \textit{and}\\
&\mathbb{E}_{(\Xi_{k,g},\Xi_{k,f}^+)}\left[\exp\left\{\lambda\big(E_k^+ - \mathbb{E}_{(\Xi_{k,g},\Xi_{k,f}^+)}[E_k^+|\mathcal{F}_k]\big)\right\}\Big|\mathcal{F}_k\right] \le \exp\left(\frac{\lambda^2\nu^2}{2}\right)\\
&\qquad\qquad \textit{for all} \;\; \lambda \in \left[0, \frac{1}{b}\right].
\end{aligned} \tag{15}$$

Assumption 3.2 (Stochastic first-order oracle)**.** *There exists $(\epsilon_g, \kappa) \in \mathbb{R}^2_{>0}$ and $\delta \in [0, 1/2)$ such that, for all $k \in \mathbb{N}$, the following inequality holds:*

$$\begin{aligned}
&\mathbb{P}_{\Xi_{k,g}}\Big[\|G_k - \nabla f(X_k)\|_2\\
&\qquad \le \max\left\{\epsilon_g, \kappa\sqrt{\|S_k\|_2^2 + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}\right\}\Big|\mathcal{F}_k\Big] \ge 1 - \delta.
\end{aligned} \tag{16}$$

Remark 3.2. *The following comments concern Assumption 3.1 and Assumption 3.2.*

1. *The latter conditions in* (14) *and* (15) *are one-sided subexponential-like assumptions on E_k and E_k^+, respectively. For more context, let us first recall the definition of sub-Gaussian and subexponential random variables. A random variable X is called sub-Gaussian if its moment-generating function satisfies $\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \le \exp(\nu^2\lambda^2/2)$ for all $\lambda \in \mathbb{R}$ for some $\nu > 0$. Such a random variable has light tails that decay at least as fast as those of a Gaussian. Subexponential random variables relax this by requiring the moment-generating function bound to hold only for $|\lambda| \le 1/b$ for some $b > 0$, thereby accommodating heavier tails. A one-sided subexponential-like random variable further relaxes this requirement by imposing the bound only for $\lambda \in [0, 1/b]$.*

2. *The conditions in* (15) *are in some sense stronger than those in* (14) *since the bounds are assumed to hold for the estimation error* E_k^+*, which is realized after the computation of a trial step* S_k*. Thus, they relate to estimation errors at the trial iterate* X_k^+*, the value of which depends on* $\mathcal{F}_k$*-measurable* X_k*, the normal step* V_k*, and the tangential step* U_k *to produce* $X_k^+ = X_k + S_k = X_k + V_k + U_k$*.*

3. *Similarly, with respect to* (16)*, a realization of* S_k *is only known after a realization of* G_k *is known. Therefore, the accuracy condition* (16) *is a posterior requirement that couples the gradient estimation error with the trial step.*

4. *The two quantities inside the max of* (16) *serve distinct roles. The constant* ϵ_g *bounds irreducible noise that is assumed to be intrinsic to the oracle. The second term in the max couples the required accuracy to the norm of the trial step* $\|S_k\|_2$*, which reflects progress towards stationarity, and to the reduction in linearized infeasibility* $\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2$*, which reflects progress towards feasibility. Together, they permit larger gradient estimation errors when significant progress is possible in either stationarity or feasibility, and demand higher accuracy when the potential progress is small.*

We also make the following assumption, which depends on the parameters defined in Assumptions 3.1 and Assumption 3.2, where the quantities u_f and p will be used in the stopping-time analysis in the next subsection.

Assumption 3.3. *With respect to the algorithmic parameter* $\bar{\epsilon}_f \in \mathbb{R}_{>0}$*, the real number*

$$u_f := \bar{\epsilon}_f - \epsilon_f$$

satisfies, with (ν, b) *from Assumption 3.1 and* δ *from Assumption 3.2,*

$$u_f \geq 0 \quad \textit{and} \quad p := 1 - \delta - \exp\left(-\min\left\{\frac{u_f^2}{2\nu^2}, \frac{u_f}{2b}\right\}\right) \in (1/2, 1). \tag{17}$$

Note that $u_f \geq 0$ whenever $\bar{\epsilon}_f \geq \epsilon_f$, i.e., when the noise level $\bar{\epsilon}_f$ used in the line search of Algorithm 1 is at least as large as the actual noise bound ϵ_f in (14). At the same time, the condition $p \in (1/2, 1]$ in (17) will hold if u_f is sufficiently large relative to ν and b and if δ is sufficiently close to zero.

Our analysis of Algorithm 1 uses a classification of the iterations based on the quality of the estimates g_k, f_k, and f_k^+ and whether the Armijo-type inequality (see Line 11 in Algorithm 1) holds. We call an iteration *true* if the estimates are sufficiently accurate, and we call an iteration *successful* if the trial step satisfies the Armijo-type inequality. We formalize these definitions next.

Definition 3.3 (True iteration)**.** *Iteration* $k \in \mathbb{N}$ *is* true *if and only if*

$$\|G_k - \nabla f(X_k)\|_2 \leq \max\left\{\epsilon_g, \kappa\sqrt{\|S_k\|_2^2 + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}\right\}$$
$$\textit{and} \quad E_k + E_k^+ \leq 2\bar{\epsilon}_f.$$

Definition 3.4 (Successful iteration)**.** *Iteration* $k \in \mathbb{N}$ *is* successful *if and only if the inequality in Line 11 of Algorithm 1 holds.*

We can now define the $\mathcal{F}_{k+1}$-measurable random variables

$$I_k := \begin{cases} 1 & \text{if iteration } k \text{ is true,} \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad \Theta_k := \begin{cases} 1 & \text{if iteration } k \text{ is successful,} \\ 0 & \text{otherwise,} \end{cases}$$

and realizations of these random variables will be denoted as i_k and θ_k, respectively.

The above classifications are used to prove a relaxed decrease condition on the merit function for iterations that are *true* and *successful* (see Lemma 3.12). The proof uses the $\mathcal{F}_{k+1/3}$-measurable random variable

$$W_k := \mathcal{T}_k(f(X_k) - f_{\text{inf}} + r(X_k) - r_{\text{inf}}) + \|c(X_k)\|_2,$$

where a realization of W_k will be denoted by w_k. The random variable W_k is a shifted version of the merit function in (11), which combines objective function gaps (to their independent infima) with the violation of the equality constraint.

We now introduce a constraint qualification assumption. The one that we adopt uses the active set at the Cauchy point for the normal step subproblem defined earlier. Specifically, it uses the $\mathcal{F}_k$-measurable random set

$$\mathcal{A}_k^c := \{i \in [n] : [X_k + V_k^c]_i = 0\},$$

which is the index set of variables active at $X_k + V_k^c$ with respect to $x \in \Omega$.

Assumption 3.4. *There exists $\sigma_{\min} \in \mathbb{R}_{>0}$ such that $\sigma_{\min}([\nabla c(X_k), I_{\mathcal{A}_k^c}^T]^T) \geq \sigma_{\min}$ for all $k \in \mathbb{N}$, where $I_{\mathcal{A}_k^c}$ denotes the subset of rows of the $n \times n$ identity matrix that correspond to the elements in $\mathcal{A}_k^c$.*

Finally, we assume that the stochastic gradient estimates are bounded uniformly. This assumption holds, for example, if f is defined as a sample average approximation. It is critical for proving that T_k is bounded away from zero.

Assumption 3.5. *There exists $\kappa_G \in \mathbb{R}_{>0}$ such that $\|G_k\|_2 \leq \kappa_G$ for all $k \in \mathbb{N}$.*

3.2 Stopping Time Analysis

Our aim in this subsection is to analyze a stopping time of our proposed algorithm. The stopping time that we consider is defined in terms of quantities introduced in the previous subsection along with a uniform lower bound for the merit parameter. Due to this latter dependence, before defining our stopping time, we prove a few introductory lemmas to show that such a lower bound exists.

We start by providing a lower bound on the reduction in linearized infeasibility.

Lemma 3.1. *There exists $\kappa_1 \in \mathbb{R}_{>0}$ such that, for all $k \in \mathbb{N}$, one has*

$$\begin{aligned}
&\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T S_k\|_2 \\
&= \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2 \\
&\geq \kappa_1 \sigma_{\min}^2 \|c(X_k)\|_2 \min\left\{\frac{1}{1 + \|\nabla c(X_k)\nabla c(X_k)^T\|_2}, \kappa_v A_k\right\} \\
&\geq \frac{\kappa_1}{\kappa_c} \sigma_{\min}^2 \|c(X_k)\|_2^2 \min\left\{\frac{1}{1 + \|\nabla c(X_k)\nabla c(X_k)^T\|_2}, \kappa_v A_k\right\}.
\end{aligned}$$

Proof. The proof follows exactly as in [7, Lemma 5.10] and uses Assumption 3.4. □ □

The next lemma bounds Δ_k in terms of the constraint violation.

Lemma 3.2. *For all $k \in \mathbb{N}$, one has $\Delta_k \leq 2\kappa_J \|c(X_k)\|_2 \leq 2\kappa_J \kappa_c$.*

Proof. See [7, Lemma 5.3], (2), and the definition of δ_k (so Δ_k) in (5). □ □

We next prove that the merit parameters are uniformly bounded from zero.

Lemma 3.3. *For all $k \in \mathbb{N}$, it holds that*

$$\mathcal{T}_{k,trial} \geq \mathcal{T}_k \geq \tau_{\min} := \min\left\{\tau_{-1}, (1 - \epsilon_\tau)\frac{(1 - 2\eta_1)\kappa_1 \sigma_{\min}^2 \min\left\{\frac{1}{1+\kappa_J^2}, \kappa_v\right\}}{2(\kappa_G + \kappa_{\partial r})\kappa_v \kappa_J + 2\kappa_v^2 \kappa_J^2 \kappa_c}\right\} > 0.$$

Proof. Let us begin by bounding the change in q_k with respect to the step S_k. Since r is convex, it follows from the subgradient inequality at $X_k + V_k$ that

$$r(X_k + V_k) - r(X_k) \leq (G_{r,k}^v)^T V_k \quad \text{for all} \quad G_{r,k}^v \in \partial r(X_k + V_k). \tag{18}$$

Since $S_k = V_k + U_k$, U_k is optimal for (7), and the zero vector is feasible for (7),

$$\begin{aligned}
&G_k^T S_k + \tfrac{1}{2A_k}\|S_k\|_2^2 + r(X_k + S_k) \\
&= G_k^T(V_k + U_k) + \tfrac{1}{2A_k}\|V_k + U_k\|_2^2 + r(X_k + V_k + U_k) \\
&= q_k(V_k + U_k) \leq q_k(V_k) = G_k^T V_k + \tfrac{1}{2A_k}\|V_k\|_2^2 + r(X_k + V_k).
\end{aligned}$$

Subtracting $q_k(0) = r(X_k)$ from both sides and using (18) gives

$$\begin{aligned}
&G_k^T S_k + \tfrac{1}{2A_k}\|S_k\|_2^2 + r(X_k + S_k) - r(X_k) \\
&\leq G_k^T V_k + \tfrac{1}{2A_k}\|V_k\|_2^2 + r(X_k + V_k) - r(X_k) \\
&\leq G_k^T V_k + \tfrac{1}{2A_k}\|V_k\|_2^2 + (G_{r,k}^v)^T V_k \\
&\overset{(i)}{\leq} (\|G_k\|_2 + \|G_{r,k}^v\|_2)\|V_k\|_2 + \tfrac{1}{2A_k}\|V_k\|_2^2 \\
&\overset{(ii)}{\leq} (\|G_k\|_2 + \|G_{r,k}^v\|_2)\kappa_v A_k \Delta_k + \tfrac{1}{2A_k}\kappa_v^2 A_k^2 \Delta_k^2 \\
&\overset{(iii)}{=} (\|G_k\|_2 + \|G_{r,k}^v\|_2)\kappa_v A_k \Delta_k + \tfrac{1}{2}\kappa_v^2 A_k \Delta_k^2 \\
&\overset{(iv)}{\leq} (\|G_k\|_2 + \|G_{r,k}^v\|_2) 2\kappa_v A_k \kappa_J \|c(X_k)\|_2 + 2\kappa_v^2 A_k \kappa_J^2 \kappa_c \|c(X_k)\|_2 \\
&\overset{(v)}{\leq} 2(\kappa_G + \kappa_{\partial r})\kappa_v \kappa_J A_k \|c(X_k)\|_2 + 2\kappa_v^2 \kappa_J^2 \kappa_c A_k \|c(X_k)\|_2,
\end{aligned} \tag{19}$$

where (i) follows from the Cauchy-Schwarz inequality, (ii) follows from the first inequality in (6), (iii) follows from canceling an A_k from the second term, (iv) follows from Lemma 3.2 and (2), and (v) follows from (2) and Assumption 3.5.

We now prove that $\{\mathcal{T}_{k,\text{trial}}\}_{k\geq 0}$ is uniformly bounded away from zero. Conditioned on $q_k(S_k) - q_k(0) \leq 0$, one finds $\mathcal{T}_{k,\text{trial}} = \infty$. Otherwise, conditioned on $q_k(S_k) - q_k(0) > 0$, we first show that $c(X_k) \neq 0$. To see this, note that if $c(X_k) = 0$ then it would follow from (5) that $\Delta_k = 0$, and then $V_k = 0$ follows from (6), and then $q_k(S_k) = q_k(V_k + U_k) \leq q_k(0)$ follows from how U_k is computed in (7), which is a contradiction. Thus, conditioned on $q_k(S_k) - q_k(0) > 0$, it must hold that $c(X_k) \neq 0$, which can be combined with the definition of $\mathcal{T}_{k,\text{trial}}$, $S_k = V_k + U_k$, $\nabla c(X_k)^T U_k = 0$ (see (8b)), Lemma 3.1, (19), (2), and the fact that $A_k \leq 1$ to get

$$\begin{aligned}
\mathcal{T}_{k,\text{trial}} &= \frac{(1-2\eta_1)(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T S_k\|_2)}{q_k(S_k) - q_k(0)} \\
&\geq \frac{(1-2\eta_1)\kappa_1 \sigma_{\min}^2 \|c(X_k)\|_2 \min\left\{\frac{1}{1+\|\nabla c(X_k)\nabla c(X_k)^T\|_2}, \kappa_v A_k\right\}}{2(\kappa_G + \kappa_{\partial r})\kappa_v \kappa_J A_k \|c(X_k)\|_2 + 2\kappa_v^2 \kappa_J^2 \kappa_c A_k \|c(X_k)\|_2} \\
&\geq \frac{(1-2\eta_1)\kappa_1 \sigma_{\min}^2 \min\left\{\frac{1}{1+\kappa_J^2}, \kappa_v A_k\right\}}{2(\kappa_G + \kappa_{\partial r})\kappa_v \kappa_J A_k + 2\kappa_v^2 \kappa_J^2 \kappa_c A_k} \\
&\geq \frac{(1-2\eta_1)\kappa_1 \sigma_{\min}^2 \min\left\{\frac{1}{1+\kappa_J^2}, \kappa_v\right\}}{2(\kappa_G + \kappa_{\partial r})\kappa_v \kappa_J + 2\kappa_v^2 \kappa_J^2 \kappa_c} > 0.
\end{aligned} \tag{20}$$

Combining the merit parameter update rule (10) and (20) completes the proof. □ □

We now define a stopping time.

Definition 3.5 (Stopping time)**.** *Choose any* $\eta \in \left(0, \frac{\sqrt{2\eta_1(1-3\eta_1)}}{1+2\sqrt{2\eta_1(1-3\eta_1)}}\right) \subset (0, \frac{1}{2})$*, and then define (see the discussion below showing* $\omega_2 > 0$*) the quantities*

$$\begin{aligned}
\omega_1 &:= \frac{1}{1-\eta_1}\left(\frac{\tau_{\min}(L_g + 2\eta_1) + L_J}{\tau_{\min}} + \frac{\kappa^2}{2\eta_1}\right) > 0, \\
\omega_2 &:= \frac{\tau_{\min} L_g + L_J}{\left(1 - \frac{\eta^2}{2\eta_1(1-2\eta)^2} - 3\eta_1\right)\tau_{\min}} > 0, \\
\kappa_T &:= \sqrt{\kappa_c}/\sqrt{\kappa_1 \sigma_{\min}^2 \min\{1/(1+\kappa_J^2), \kappa_v\}} > 0, \quad \textit{and} \\
\bar{\alpha} &:= \gamma^{\bar{i}+1}\alpha_0 \quad \textit{with } \bar{i} \textit{ defined as the smallest nonnegative integer} \\
&\qquad\qquad \textit{such that } \gamma^{\bar{i}}\alpha_0 \le \min\{\omega_1^{-1}, \omega_2^{-1}\},
\end{aligned}$$

where we note that this ensures that $\bar{\alpha} \le \gamma\alpha_0 < \alpha_0$*. Then, for* $\theta \in (0,1)$ *and any*

$$\epsilon > \max\{\sqrt{2}, \kappa_T\} \max\left\{\tfrac{\epsilon_g}{\eta}, \max\left\{1+\kappa, \tfrac{1}{\sqrt{1-2\eta}}\right\}\sqrt{\tfrac{(4+4\tau_{-1})\bar{\epsilon}_f}{\theta\tau_{\min}\bar{\alpha}\eta_1(p-1/2)}}\right\} \tag{21}$$

let the stopping time be the random variable

$$T_\epsilon := \min\left\{k \in \mathbb{N} : \Psi_k \le \frac{\epsilon}{\max\{\sqrt{2}, \kappa_T\}}\right\}, \tag{22}$$

where, for each $k \in \mathbb{N}$*, the optimality measure is defined as*

$$\Psi_k := \sqrt{\frac{\|S_k^{\text{true}}\|_2^2}{2A_k^2} + \frac{\|\min\{X_k, -Z_k^{\text{true}}\}\|_2^2}{2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}}.$$

We allow $T_\epsilon = \infty$ *whenever the minimum in* (22) *is not attained.*

In Definition 3.5, it is straightforward to see that $\omega_1 > 0$ and $\kappa_T > 0$, and once it is shown that $\omega_2 > 0$ it is also straightforward to see that $\bar{\alpha} > 0$. To show $\omega_2 > 0$, first let $\bar{a} := \sqrt{2\eta_1(1-3\eta_1)} > 0$ so that we may rewrite the interval containing η as $\eta \in (0, \frac{\bar{a}}{1+2\bar{a}})$. To prove that $\omega_2 > 0$, it is sufficient to show that

$$\frac{\eta^2}{2\eta_1(1-2\eta)^2} < 1 - 3\eta_1 = \frac{\bar{a}^2}{2\eta_1}.$$

Multiplying both sides by $2\eta_1(1-2\eta)^2 > 0$ yields

$$\eta^2 < \bar{a}^2(1-2\eta)^2 \iff \eta < \bar{a}(1-2\eta) \iff (1+2\bar{a})\eta < \bar{a} \iff \eta < \frac{\bar{a}}{1+2\bar{a}},$$

which holds by the definition of η, thus confirming that $\omega_2 > 0$.

We now justify the definition of the stopping time T_ϵ. To do this, the next lemma shows that, for a given run of Algorithm 1, an iterate satisfying the stopping time condition is an ϵ-KKT point, which we define first.

Definition 3.6. *The kth iterate of Algorithm 1 is an* ϵ*-KKT point if and only if*

$$\max\left\{\|\nabla f(x_k) + g_{r,k}^{true} + \nabla c(x_k) y_k^{true} + z_k^{true}\|_2, \|\min\{x_k, -z_k^{true}\}\|_2, \|c(x_k)\|_2\right\} \le \epsilon$$

where we also remind the reader that $g_{r,k}^{true} \in \partial r(x_k + s_k^{true})$*.*

Lemma 3.4. *If during iteration $\bar{k} \in \mathbb{N}$ the tuple $(x_{\bar{k}}, \alpha_{\bar{k}}, v_{\bar{k}}, u_{\bar{k}}^{true}, g_{r,\bar{k}}^{true}, y_{\bar{k}}^{true}, z_{\bar{k}}^{true}, s_{\bar{k}}^{true})$ satisfies the inequality defining T_ϵ, then $x_{\bar{k}}$ is an ϵ-KKT point.*

Proof. Consider such a $\bar{k}$. It follows from Definition 3.1, the fact that the stopping time condition ensures $\|s_{\bar{k}}^{\text{true}}/(\sqrt{2}\alpha_{\bar{k}})\|_2 \leq \epsilon/\max\{\sqrt{2}, \kappa_T\}$ (and similarly for the other terms in the left-hand side of the stopping-time condition), Lemma 3.1, (2), $A_k \leq 1$, and Definition 3.5 that

$$\begin{aligned} \|\nabla f(x_{\bar{k}}) + g_{r,\bar{k}}^{\text{true}} + \nabla c(x_{\bar{k}}) y_{\bar{k}}^{\text{true}} + z_{\bar{k}}^{\text{true}}\|_2 = \left\|s_{\bar{k}}^{\text{true}}/\alpha_{\bar{k}}\right\|_2 &\leq \tfrac{\sqrt{2}\epsilon}{\max\{\sqrt{2},\kappa_T\}} \leq \epsilon, \\ \|\min\{x_{\bar{k}}, -z_{\bar{k}}^{\text{true}}\}\|_2 &\leq \tfrac{\sqrt{2}\epsilon}{\max\{\sqrt{2},\kappa_T\}} \leq \epsilon, \quad \text{and} \\ \|c(x_{\bar{k}})\|_2 &\leq \tfrac{\kappa_T\epsilon}{\max\{\sqrt{2},\kappa_T\}} \leq \epsilon. \end{aligned}$$

It follows from these inequalities and Definition 3.6 that $x_{\bar{k}}$ is an ϵ-KKT point. □ □

Before commencing the main part of our analysis, we motivate the definition of $\bar{\alpha}$ in Definition 3.5 by proving the following technical lemma and subsequently stating a definition for "large" proximal-parameters.

Lemma 3.5. *For all $k \in \mathbb{N}$, $\min\{A_k, A_{k+1}\} < \bar{\alpha}$ if and only if $\max\{A_k, A_{k+1}\} \leq \bar{\alpha}$.*

Proof. Observe that Definition 3.5 ensures that $\bar{\alpha} \leq \gamma\alpha_0$. This fact, along with the updating strategy for the proximal parameter, ensure that if either $\min\{A_k, A_{k+1}\} < \bar{\alpha}$ or $\max\{A_k, A_{k+1}\} \leq \bar{\alpha}$, then

$$\max\{A_k, A_{k+1}\} = \gamma^{-1}\min\{A_k, A_{k+1}\} > \min\{A_k, A_{k+1}\}. \tag{23}$$

Thus, if $\max\{A_k, A_{k+1}\} \leq \bar{\alpha}$, then it follows from (23) that $\min\{A_k, A_{k+1}\} < \max\{A_k, A_{k+1}\} \leq \bar{\alpha}$. Otherwise, if $\min\{A_k, A_{k+1}\} < \bar{\alpha}$, then it follows from how $\bar{\alpha}$ is defined that $\gamma^{-1}\min\{A_k, A_{k+1}\} \leq \bar{\alpha}$. Combining this inequality with the equation in (23) shows that $\max\{A_k, A_{k+1}\} \leq \bar{\alpha}$, which completes the proof. □ □

The lemma above can now be used to define the following random variable.

Definition 3.7 (Large proximal parameters)**.** *For each $k \in \mathbb{N}$, let*

$$L_k := \begin{cases} 1, & \text{if } \min\{A_k, A_{k+1}\} \geq \bar{\alpha}, \\ 0, & \text{if } \max\{A_k, A_{k+1}\} \leq \bar{\alpha}, \end{cases}$$

which is well defined because of Lemma 3.5. We also note that L_k is $\mathcal{F}_{k+1}$-measurable since A_k is $\mathcal{F}_k$-measurable and Θ_k is $\mathcal{F}_{k+1}$-measurable.

We now start the main part of our stopping-time analysis. The next result bounds the difference between S_k and S_k^{true}, and between U_k and U_k^{true}.

Lemma 3.6. *For $k \in \mathbb{N}$, $\frac{1}{A_k}\|S_k - S_k^{\text{true}}\|_2 = \frac{1}{A_k}\|U_k - U_k^{\text{true}}\|_2 \leq \|G_k - \nabla f(X_k)\|_2$.*

Proof. The claimed equality follows from $\|S_k - S_k^{\text{true}}\|_2 = \|V_k + U_k - (V_k + U_k^{\text{true}})\|_2 = \|U_k - U_k^{\text{true}}\|_2$. The claimed inequality follows trivially if $\|U_k - U_k^{\text{true}}\|_2 = 0$, so for the remainder of the proof we assume that $\|U_k - U_k^{\text{true}}\|_2 \neq 0$. Consider arbitrary $k \in \mathbb{N}$ and define $h_k : \mathbb{R}^n \to \mathbb{R}$ by

$$h_k(u) = r(X_k + V_k + u) + I_\Omega(X_k + V_k + u),$$

where I_Ω denotes the indicator function associated with the set Ω. Note that h_k is convex since both r and I_Ω are convex. One can now write U_k and U_k^{true}, respectively, as the unique solutions to the following subproblems:

$$\begin{aligned} U_k^{\text{true}} = \arg\min_{u\in\mathbb{R}^n} \ & \nabla f(X_k)^T u + \tfrac{1}{2A_k}\|u\|_2^2 + \tfrac{1}{A_k}V_k^T u + h_k(u) \\ \text{subject to } & \nabla c(X_k)^T u = 0 \quad \text{and} \end{aligned}$$

$$U_k = \arg\min_{u\in\mathbb{R}^n} \; G_k^T u + \tfrac{1}{2A_k}\|u\|_2^2 + \tfrac{1}{A_k} V_k^T u + h_k(u)$$
$$\text{subject to } \nabla c(X_k)^T u = 0.$$

Using the subdifferential sum rule [20], which holds since the constraint qualification $\mathrm{dom}(r)\cap\mathrm{int}\big(\mathrm{dom}(I_\Omega)\big) \neq \emptyset$ holds trivially since $\mathrm{dom}(r) = \mathbb{R}^n$, it follows that for all $u \in \mathbb{R}^n$ one has

$$\partial h_k(u) = \partial r(X_k + V_k + u) + N_\Omega(X_k + V_k + u).$$

Here, $N_\Omega(X_k+V_k+u)$ is the normal cone to Ω at X_k+V_k+u. From the KKT conditions of the optimization problems above, there exist $(G_{h,k}^{\text{true}}, G_{h,k}) \in \partial h_k(U_k^{\text{true}}) \times \partial h_k(U_k)$ and $(Y_k^{\text{true}}, Y_k) \in \mathbb{R}^m \times \mathbb{R}^m$ such that

$$\nabla f(X_k) + \tfrac{1}{A_k}(U_k^{\text{true}} + V_k) + G_{h,k}^{\text{true}} + \nabla c(X_k) Y_k^{\text{true}} = 0 \;\text{ and}$$
$$G_k + \tfrac{1}{A_k}(U_k + V_k) + G_{h,k} + \nabla c(X_k) Y_k = 0.$$

Subtracting these two equations gives

$$(G_k - \nabla f(X_k)) + \tfrac{1}{A_k}(U_k - U_k^{\text{true}}) + (G_{h,k} - G_{h,k}^{\text{true}}) + \nabla c(X_k)(Y_k - Y_k^{\text{true}}) = 0.$$

Multiplying by $(U_k - U_k^{\text{true}})^T$ and using $\nabla c(X_k)^T U_k^{\text{true}} = \nabla c(X_k)^T U_k = 0$ gives

$$(G_k - \nabla f(X_k))^T (U_k - U_k^{\text{true}}) + \tfrac{1}{A_k}\|U_k - U_k^{\text{true}}\|_2^2 + (G_{h,k} - G_{h,k}^{\text{true}})^T (U_k - U_k^{\text{true}}) = 0.$$

Since h_k is convex, its subdifferential ∂h_k is a maximal monotone operator, from which it follows that $(G_{h,k} - G_{h,k}^{\text{true}})^T (U_k - U_k^{\text{true}}) \geq 0$. Combining this with the equation above and using the Cauchy-Schwarz inequality gives

$$\begin{aligned}\tfrac{1}{A_k}\|U_k - U_k^{\text{true}}\|_2^2 &\leq -(G_k - \nabla f(X_k))^T (U_k - U_k^{\text{true}}) \\ &\leq \|G_k - \nabla f(X_k)\|_2 \|U_k - U_k^{\text{true}}\|_2.\end{aligned}$$

Dividing both sides by $\|U_k - U_k^{\text{true}}\|_2 \neq 0$ completes the proof. □ □

We now prove an upper bound on the change in the merit function attained by the trial step S_k in all iterations.

Lemma 3.7. *For all $k \in \mathbb{N}$, it holds that*

$$\begin{aligned}&\phi(X_k + S_k, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k) \\ \leq\; & -\tfrac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + \tfrac{\mathcal{T}_k L_g + L_J}{2}\|S_k\|_2^2 + \mathcal{T}_k(\nabla f(X_k) - G_k)^T S_k + \mathcal{T}_k G_k^T S_k \\ & + \mathcal{T}_k(r(X_k + S_k) - r(X_k)) + \tfrac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + \|c(X_k) + \nabla c(X_k)^T V_k\|_2 - \|c(X_k)\|_2.\end{aligned}$$

Proof. Consider arbitrary $k \in \mathbb{N}$. Lipschitz continuity of $\nabla f(\cdot)$ and $\nabla c(\cdot)^T$ (see Assumption 1.1), $S_k = V_k + U_k$, and $\nabla c(X_k)^T U_k = 0$ show that

$$\begin{aligned}&\phi(X_k + S_k, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k) \\ =\; & \mathcal{T}_k(f(X_k + S_k) + r(X_k + S_k)) + \|c(X_k + S_k)\|_2 \\ & - (\mathcal{T}_k(f(X_k) + r(X_k)) + \|c(X_k)\|_2) \\ \leq\; & \mathcal{T}_k(\nabla f(X_k)^T S_k + \tfrac{L_g}{2}\|S_k\|_2^2) + \mathcal{T}_k(r(X_k + S_k) - r(X_k)) \\ & + \|c(X_k) + \nabla c(X_k)^T S_k\|_2 - \|c(X_k)\|_2 + \tfrac{L_J}{2}\|S_k\|_2^2 \\ =\; & -\tfrac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + \tfrac{\mathcal{T}_k L_g + L_J}{2}\|S_k\|_2^2 + \mathcal{T}_k(\nabla f(X_k) - G_k)^T S_k + \mathcal{T}_k G_k^T S_k \\ & + \mathcal{T}_k(r(X_k + S_k) - r(X_k)) + \tfrac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + \|c(X_k) + \nabla c(X_k)^T V_k\|_2 - \|c(X_k)\|_2,\end{aligned}$$

which completes the proof. □ □

Lemma 3.7 provides an upper bound on the change in the merit function. The next lemma bounds the term $\mathcal{T}_k(\nabla f(X_k) - G_k)^T S_k$ appearing in this upper bound in terms of measures related to stationarity.

Lemma 3.8. *Conditioned on $k < T_\epsilon$, the following hold.*

(i) If $\|G_k - \nabla f(X_k)\|_2 \le \kappa\sqrt{\|S_k\|_2^2 + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}$, then

$$\begin{aligned}\mathcal{T}_k(\nabla f(X_k) - G_k)^T S_k \le\ & \left(\tfrac{\kappa^2}{4\eta_1} + \eta_1\right)\mathcal{T}_k A_k \tfrac{\|S_k\|_2^2}{A_k} \\ & + \eta_1(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2).\end{aligned}$$

(ii) If $\|G_k - \nabla f(X_k)\|_2 \le \epsilon_g$, then

$$\begin{aligned}\mathcal{T}_k(\nabla f(X_k) - G_k)^T S_k \le\ & \left(\tfrac{\eta^2}{4\eta_1(1-2\eta)^2} + \eta_1\right)\mathcal{T}_k \tfrac{\|S_k\|_2^2}{A_k} \\ & + \eta_1(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2).\end{aligned}$$

Proof. Condition on the event that $k < T_\epsilon$. First, suppose that the condition in (i) holds. Then, along with the Cauchy-Schwarz inequality and the fact that $\nabla c(X_k)^T U_k = 0$ (see (8b)) shows that

$$\begin{aligned}&\mathcal{T}_k(\nabla f(X_k) - G_k)^T S_k \\ &\le \mathcal{T}_k\|\nabla f(X_k) - G_k\|_2\|S_k\|_2 \\ &\le \kappa\mathcal{T}_k\sqrt{\|S_k\|_2^2 + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}\|S_k\|_2. \end{aligned} \tag{24}$$

Now $\mathcal{T}_k \le 1$ and Young's inequality with $\delta_Y \in \mathbb{R}_{>0}$ (also known as the Peter-Paul inequality), which is that $ab \le \frac{a^2}{2\delta_Y} + \frac{\delta_Y b^2}{2}$ for any $(a,b) \in \mathbb{R}_{\ge 0} \times \mathbb{R}_{\ge 0}$, shows with $\delta_Y = \frac{2\eta_1}{\kappa}$, $a = \|S_k\|_2$, and $b = \sqrt{\|S_k\|_2^2 + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}$ that

$$\begin{aligned}&\kappa\mathcal{T}_k\|S_k\|_2\sqrt{\|S_k\|_2^2 + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2} \\ &\le \kappa\mathcal{T}_k\left(\tfrac{\|S_k\|_2^2}{2\delta_Y} + \tfrac{\delta_Y}{2}\left(\|S_k\|_2^2 + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2\right)\right) \\ &= \kappa\mathcal{T}_k\left(\tfrac{1}{2\delta_Y} + \tfrac{\delta_Y}{2}\right)\|S_k\|_2^2 + \tfrac{\kappa\mathcal{T}_k\,\delta_Y}{2}(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2) \\ &= \left(\tfrac{\kappa^2}{4\eta_1} + \eta_1\right)\mathcal{T}_k\|S_k\|_2^2 + \eta_1\mathcal{T}_k(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2) \\ &\le \left(\tfrac{\kappa^2}{4\eta_1} + \eta_1\right)\mathcal{T}_k A_k\tfrac{\|S_k\|_2^2}{A_k} + \eta_1\left(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2\right).\end{aligned}$$

Combining this inequality with (24) completes the proof of part (i).

On the other hand, suppose the condition in (ii) holds. From $A_k \le 1$, $\|S_k^{\text{true}}\|_2 \ge \|\min\{X_k, -Z_k^{\text{true}}\}\|_2$ (see [7, Lemma 4.6]), and (21), one obtains

$$\sqrt{\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}} > \frac{\epsilon}{\max\{\sqrt{2}, \kappa_T\}} > \frac{\epsilon_g}{\eta}, \tag{25}$$

which implies along with the condition in (ii) that

$$\|G_k - \nabla f(X_k)\|_2 \le \epsilon_g < \eta\sqrt{\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}}. \tag{26}$$

Thus, with the Cauchy-Schwarz inequality, Lemma 3.6, and $\nabla c(X_k)^T U_k = 0$,

$$\left(\frac{\|S_k^{\text{true}}\|_2^2}{A_k} + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2\right)$$

$$
\begin{aligned}
&- \left(\frac{\|S_k\|_2^2}{A_k} + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2 \right) \\
&= \frac{\|S_k^{\text{true}}\|_2^2 - \|S_k\|_2^2}{A_k} \\
&= \frac{1}{A_k} \left(2(S_k^{\text{true}})^T (S_k^{\text{true}} - S_k) - \|S_k^{\text{true}} - S_k\|_2^2 \right) \\
&\leq \frac{2\|S_k^{\text{true}}\|_2 \|S_k^{\text{true}} - S_k\|_2}{A_k} \\
&\leq 2\|S_k^{\text{true}}\|_2 \|G_k - \nabla f(X_k)\|_2 \\
&\leq 2\eta \|S_k^{\text{true}}\|_2 \sqrt{\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}} \\
&\leq 2\eta \cdot A_k \sqrt{\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}} \\
&\qquad \cdot \sqrt{\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}} \\
&= 2\eta A_k \left(\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k} \right) \\
&= 2\eta \left(\frac{\|S_k^{\text{true}}\|_2^2}{A_k} + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2 \right),
\end{aligned}
$$

which, after rearrangement, gives

$$
\begin{aligned}
&\frac{\|S_k^{\text{true}}\|_2^2}{A_k} + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2 \\
&\leq \frac{1}{1-2\eta} \left(\frac{\|S_k\|_2^2}{A_k} + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2 \right).
\end{aligned}
$$

Dividing both sides by A_k and taking square roots yields

$$
\begin{aligned}
&\sqrt{\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}} \\
&\leq \frac{1}{\sqrt{1-2\eta}} \sqrt{\frac{\|S_k\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}}.
\end{aligned} \tag{27}
$$

Combining the Cauchy–Schwarz inequality, (26), (27), and the fact that $1 - 2\eta \in (0, 1)$ (since $\eta \in (0, \frac{1}{2})$ under Definition 3.5), one finds that

$$
\begin{aligned}
&\mathcal{T}_k (\nabla f(X_k) - G_k)^T S_k \\
&\leq \mathcal{T}_k \|\nabla f(X_k) - G_k\|_2 \|S_k\|_2 \\
&\leq \eta \mathcal{T}_k \|S_k\|_2 \sqrt{\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}} \\
&\leq \frac{\eta \mathcal{T}_k}{\sqrt{1-2\eta}} \|S_k\|_2 \sqrt{\frac{\|S_k\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}} \\
&\leq \frac{\eta \mathcal{T}_k}{1-2\eta} \|S_k\|_2 \sqrt{\frac{\|S_k\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}}.
\end{aligned} \tag{28}
$$

To bound the right-hand side of (28), one can apply Young's inequality with $\delta_Y = \frac{2A_k\eta_1(1-2\eta)}{\eta} > 0$, $a = \|S_k\|_2$, and $b = \sqrt{\frac{\|S_k\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2}{A_k}}$ and use the fact that $\mathcal{T}_k \leq 1$ to obtain

$$
\begin{aligned}
&\frac{\eta\mathcal{T}_k}{1-2\eta}\|S_k\|_2\sqrt{\frac{\|S_k\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2}{A_k}} \\
\leq\ & \frac{\eta\mathcal{T}_k}{1-2\eta}\left(\frac{\|S_k\|_2^2}{2\delta_Y} + \frac{\delta_Y}{2}\left(\frac{\|S_k\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2}{A_k}\right)\right) \\
=\ & \frac{\eta\mathcal{T}_k}{1-2\eta}\left(\frac{1}{2\delta_Y} + \frac{\delta_Y}{2A_k^2}\right)\|S_k\|_2^2 + \frac{\frac{\eta\mathcal{T}_k}{1-2\eta}\delta_Y}{2A_k}\left(\|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2\right) \\
=\ & \frac{\eta\mathcal{T}_k}{1-2\eta}\left(\frac{\frac{\eta}{1-2\eta}}{4A_k\eta_1} + \frac{A_k\eta_1}{\frac{\eta}{1-2\eta}A_k^2}\right)\|S_k\|_2^2 + \eta_1\mathcal{T}_k(\|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2) \\
\leq\ & \left(\frac{\eta^2\mathcal{T}_k}{4\eta_1(1-2\eta)^2} + \eta_1\mathcal{T}_k\right)\frac{\|S_k\|_2^2}{A_k} + \eta_1\left(\|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2\right) \\
=\ & \left(\frac{\eta^2}{4\eta_1(1-2\eta)^2} + \eta_1\right)\mathcal{T}_k\frac{\|S_k\|_2^2}{A_k} + \eta_1\left(\|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2\right).
\end{aligned}
$$

Combining this inequality with (28) completes the proof of part (ii). □ □

We now use Lemma 3.7 and Lemma 3.8 to obtain a refined upper bound on the merit function decrease in iterations before the stopping time condition holds.

Lemma 3.9. *Conditioned on $k < T_\epsilon$, the following hold.*

(i) If $\|G_k - \nabla f(X_k)\|_2 \leq \kappa\sqrt{\|S_k\|_2^2 + \|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2}$, then

$$
\begin{aligned}
&\phi(X_k+S_k,\mathcal{T}_k) - \phi(X_k,\mathcal{T}_k) \\
\leq\ & -\frac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + \frac{\mathcal{T}_k L_g + L_J}{2}\|S_k\|_2^2 + \left(\frac{\kappa^2}{4\eta_1} + \eta_1\right)A_k\mathcal{T}_k\frac{\|S_k\|_2^2}{A_k} \\
&+ \eta_1(\|c(X_k)+\nabla c(X_k)^T V_k\|_2 - \|c(X_k)\|_2).
\end{aligned}
$$

(ii) If $\|G_k - \nabla f(X_k)\|_2 \leq \epsilon_g$, then

$$
\begin{aligned}
&\phi(X_k+S_k,\mathcal{T}_k) - \phi(X_k,\mathcal{T}_k) \\
\leq\ & -\frac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + \frac{\mathcal{T}_k L_g + L_J}{2}\|S_k\|_2^2 + \left(\frac{\eta^2}{4\eta_1(1-2\eta)^2} + \eta_1\right)\mathcal{T}_k\frac{\|S_k\|_2^2}{A_k} \\
&+ \eta_1(\|c(X_k)+\nabla c(X_k)^T V_k\|_2 - \|c(X_k)\|_2).
\end{aligned}
$$

Proof. Condition on the event that $k < T_\epsilon$. Let us first establish a bound that is common to both part (i) and part (ii). In particular, let us show that

$$
\begin{aligned}
&\mathcal{T}_k(G_k^T S_k + \tfrac{1}{2A_k}\|S_k\|_2^2 + r(X_k+S_k) - r(X_k)) \\
&+ \|c(X_k)+\nabla c(X_k)^T V_k\|_2 - \|c(X_k)\|_2 \\
\leq\ & -2\eta_1(\|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2).
\end{aligned} \tag{29}
$$

Note that the term $\|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2$ in the right-hand side of (29) is nonnegative due to Lemma 3.1. Let us now consider the two cases in (9).

Case 1: Suppose that $G_k^T S_k + \frac{1}{2A_k}\|S_k\|_2^2 + r(X_k+S_k) - r(X_k) \le 0$. Since $\eta_1 \in (0, 1/3)$, one has $0 < 2\eta_1 < 1$, and therefore

$$\begin{aligned}
&\mathcal{T}_k(G_k^T S_k + \tfrac{1}{2A_k}\|S_k\|_2^2 + r(X_k+S_k) - r(X_k)) \\
&\quad + \|c(X_k) + \nabla c(X_k)^T V_k\|_2 - \|c(X_k)\|_2 \\
\le\ & -(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2) \\
\le\ & -2\eta_1(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2),
\end{aligned}$$

thus verifying that (29) holds in this case, as claimed.

Case 2: Suppose that $G_k^T S_k + \frac{1}{2A_k}\|S_k\|_2^2 + r(X_k+S_k) - r(X_k) > 0$. It follows from (9) and Lemma 3.3 that

$$\mathcal{T}_{k,\text{trial}} = \frac{(1-2\eta_1)(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T S_k\|_2)}{G_k^T S_k + \frac{1}{2A_k}\|S_k\|_2^2 + r(X_k+S_k) - r(X_k)}$$

is finite and strictly positive with $\mathcal{T}_{k,\text{trial}} \ge \mathcal{T}_k > 0$. Combining this fact with the value of $\mathcal{T}_{k,\text{trial}}$ above, $G_k^T S_k + \frac{1}{2A_k}\|S_k\|_2^2 + r(X_k+S_k) - r(X_k) > 0$, and the fact that $\|c(X_k) + \nabla c(X_k)^T S_k\|_2 = \|c(X_k) + \nabla c(X_k)^T V_k\|_2$ (see (8b)) gives

$$\begin{aligned}
&\mathcal{T}_k\big(G_k^T S_k + \tfrac{1}{2A_k}\|S_k\|_2^2 + r(X_k+S_k) - r(X_k)\big) \\
&\le (1-2\eta_1)(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2).
\end{aligned}$$

Therefore, one finds that

$$\begin{aligned}
&\mathcal{T}_k(G_k^T S_k + \tfrac{1}{2A_k}\|S_k\|_2^2 + r(X_k+S_k) - r(X_k)) \\
&\quad + \|c(X_k) + \nabla c(X_k)^T V_k\|_2 - \|c(X_k)\|_2 \\
&\le (1-2\eta_1)(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2) \\
&\quad - (\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2) \\
&= -2\eta_1(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2),
\end{aligned}$$

again verifying that (29) holds. Overall, (29) holds in all cases, as claimed.

Now suppose the condition in (i) holds. Lemma 3.7 gives

$$\begin{aligned}
&\phi(X_k+S_k, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k) \\
\le\ & -\tfrac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + \tfrac{\mathcal{T}_k L_g + L_J}{2}\|S_k\|_2^2 + \mathcal{T}_k(\nabla f(X_k) - G_k)^T S_k \\
&+ \mathcal{T}_k(G_k^T S_k + \tfrac{1}{2A_k}\|S_k\|_2^2 + r(X_k+S_k) - r(X_k)) \\
&+ \|c(X_k) + \nabla c(X_k)^T V_k\|_2 - \|c(X_k)\|_2.
\end{aligned}$$

Using Lemma 3.8(i) and (29) gives

$$\begin{aligned}
&\phi(X_k+S_k, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k) \\
\le\ & -\tfrac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + \tfrac{\mathcal{T}_k L_g + L_J}{2}\|S_k\|_2^2 + \left(\tfrac{\kappa^2}{4\eta_1} + \eta_1\right)\mathcal{T}_k A_k \tfrac{\|S_k\|_2^2}{A_k} \\
&+ \eta_1(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2) \\
&- 2\eta_1(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2) \\
=\ & -\tfrac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + \tfrac{\mathcal{T}_k L_g + L_J}{2}\|S_k\|_2^2 + \left(\tfrac{\kappa^2}{4\eta_1} + \eta_1\right) A_k \mathcal{T}_k \tfrac{\|S_k\|_2^2}{A_k} \\
&+ \eta_1(\|c(X_k) + \nabla c(X_k)^T V_k\|_2 - \|c(X_k)\|_2),
\end{aligned}$$

which is the bound claimed in part (i).

Now suppose the condition in (ii) holds. Lemma 3.7, Lemma 3.8(ii), and (29) give

$$\begin{aligned}
&\phi(X_k+S_k,\mathcal{T}_k)-\phi(X_k,\mathcal{T}_k)\\
\le\ & -\tfrac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2+\tfrac{\mathcal{T}_kL_g+L_J}{2}\|S_k\|_2^2+\left(\tfrac{\eta^2}{4\eta_1(1-2\eta)^2}+\eta_1\right)\mathcal{T}_k\tfrac{\|S_k\|_2^2}{A_k}\\
&+\eta_1(\|c(X_k)\|_2-\|c(X_k)+\nabla c(X_k)^TV_k\|_2)\\
&-2\eta_1(\|c(X_k)\|_2-\|c(X_k)+\nabla c(X_k)^TV_k\|_2)\\
=\ & -\tfrac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2+\tfrac{\mathcal{T}_kL_g+L_J}{2}\|S_k\|_2^2+\left(\tfrac{\eta^2}{4\eta_1(1-2\eta)^2}+\eta_1\right)\mathcal{T}_k\tfrac{\|S_k\|_2^2}{A_k}\\
&+\eta_1(\|c(X_k)+\nabla c(X_k)^TV_k\|_2-\|c(X_k)\|_2),
\end{aligned}$$

which is the bound claimed in part (ii). □ □

Our next result is needed because the stopping time is defined using S_k^{true}, whereas the algorithm only has access to S_k.

Lemma 3.10. *Conditioned on $k<T_\epsilon$, the following hold.*

(i) If $\|G_k-\nabla f(X_k)\|_2\le\kappa\sqrt{\|S_k\|_2^2+\|c(X_k)\|_2-\|c(X_k)+\nabla c(X_k)^TV_k\|_2}$, then

$$\begin{aligned}
&\left(\mathcal{T}_k\frac{\|S_k\|_2^2}{2A_k}+\|c(X_k)\|_2-\|c(X_k)+\nabla c(X_k)^TV_k\|_2\right)\\
\ge\ &\frac{1}{(1+\kappa)^2}\left(\mathcal{T}_k\frac{\|S_k^{\text{true}}\|_2^2}{2A_k}+\|c(X_k)\|_2-\|c(X_k)+\nabla c(X_k)^TV_k\|_2\right).
\end{aligned}$$

(ii) If $\|G_k-\nabla f(X_k)\|_2\le\epsilon_g$, then

$$\begin{aligned}
&\left(\mathcal{T}_k\frac{\|S_k\|_2^2}{2A_k}+\|c(X_k)\|_2-\|c(X_k)+\nabla c(X_k)^TV_k\|_2\right)\\
\ge\ &(1-2\eta)\left(\mathcal{T}_k\frac{\|S_k^{\text{true}}\|_2^2}{2A_k}+\|c(X_k)\|_2-\|c(X_k)+\nabla c(X_k)^TV_k\|_2\right).
\end{aligned}$$

Proof. Condition on the event that $k<T_\epsilon$. Suppose that the condition in (i) holds. It follows from the condition in (i), the Cauchy-Schwarz inequality, $A_k\le1$, $\mathcal{T}_k\le1$, and Lemma 3.6 that

$$\begin{aligned}
&\left(\mathcal{T}_k\frac{\|S_k^{\text{true}}\|_2^2}{2A_k}+\|c(X_k)\|_2-\|c(X_k)+\nabla c(X_k)^TV_k\|_2\right)\\
&-\left(\mathcal{T}_k\frac{\|S_k\|_2^2}{2A_k}+\|c(X_k)\|_2-\|c(X_k)+\nabla c(X_k)^TV_k\|_2\right)\\
=\ &\mathcal{T}_k\frac{\|S_k^{\text{true}}\|_2^2-\|S_k\|_2^2}{2A_k}\\
=\ &\frac{\mathcal{T}_k}{2A_k}\left((S_k^{\text{true}}-S_k)^T(S_k^{\text{true}}-S_k)+2S_k^T(S_k^{\text{true}}-S_k)\right)\\
\le\ &\frac{\mathcal{T}_k}{2}\left(\frac{\|S_k^{\text{true}}-S_k\|_2^2}{A_k^2}+2\|S_k\|_2\frac{\|S_k^{\text{true}}-S_k\|_2}{A_k}\right)\\
\le\ &\frac{\mathcal{T}_k}{2}\Bigg(\kappa^2\left(\|S_k\|_2^2+\|c(X_k)\|_2-\|c(X_k)+\nabla c(X_k)^TV_k\|_2\right)\\
&\qquad+2\|S_k\|_2\kappa\sqrt{\|S_k\|_2^2+\|c(X_k)\|_2-\|c(X_k)+\nabla c(X_k)^TV_k\|_2}\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\mathcal{T}_k(2\kappa+\kappa^2)}{2}\left(\frac{\|S_k\|_2^2}{A_k} + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2\right) \\
&\leq (2\kappa+\kappa^2)\left(\mathcal{T}_k\frac{\|S_k\|_2^2}{2A_k} + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2\right),
\end{aligned}
$$

which after rearrangement completes the proof of part (i).

Now suppose the condition in (ii) holds. Using the Cauchy-Schwarz inequality, inequality (26) in the proof of Lemma 3.8(ii), Lemma 3.6, and $\mathcal{T}_k \leq 1$ it follows that

$$
\begin{aligned}
&\left(\frac{\mathcal{T}_k\|S_k^{\text{true}}\|_2^2}{2A_k} + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2\right) \\
&\quad - \left(\frac{\mathcal{T}_k\|S_k\|_2^2}{2A_k} + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2\right) \\
&= \mathcal{T}_k\frac{\left(\|S_k^{\text{true}}\|_2^2 - \|S_k\|_2^2\right)}{2A_k} \\
&= \frac{\mathcal{T}_k}{2A_k}\left(2(S_k^{\text{true}})^T(S_k^{\text{true}} - S_k) - \|S_k^{\text{true}} - S_k\|_2^2\right) \\
&\leq \frac{\mathcal{T}_k\|S_k^{\text{true}}\|_2\|S_k^{\text{true}} - S_k\|_2}{A_k} \\
&\leq \mathcal{T}_k\|S_k^{\text{true}}\|_2\|G_k - \nabla f(X_k)\|_2 \\
&\leq \eta\mathcal{T}_k\|S_k^{\text{true}}\|_2\sqrt{\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}} \\
&\leq \eta\mathcal{T}_k A_k\sqrt{\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{2(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2)}{A_k}} \\
&\qquad \cdot\sqrt{\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{2(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2)}{A_k}} \\
&= \eta\mathcal{T}_k A_k\left(\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{2(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2)}{A_k}\right) \\
&= 2\eta\mathcal{T}_k\left(\frac{\|S_k^{\text{true}}\|_2^2}{2A_k} + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2\right) \\
&\leq 2\eta\left(\mathcal{T}_k\frac{\|S_k^{\text{true}}\|_2^2}{2A_k} + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2\right),
\end{aligned}
$$

which after rearrangement completes the proof for part (ii). □ □

The following lemma bounds some key differences between iterations.

Lemma 3.11. *For all $k \in \mathbb{N}$, it holds that*

$$
|(\Phi_k^+ - \Phi_k) - (\phi(X_k^+, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k))| \leq \mathcal{T}_k(E_k + E_k^+) \quad \textit{and} \tag{30a}
$$

$$
W_{k+1} - W_k \leq \phi(X_{k+1}, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k). \tag{30b}
$$

Proof. It follows from the definitions of Φ_k^+ and Φ_k that

$$
\Phi_k^+ - \Phi_k = \phi(X_k^+, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k) - \mathcal{T}_k(f(X_k^+) - F_k^+) + \mathcal{T}_k(f(X_k) - F_k).
$$

After rearrangement and using the triangle inequality, one obtains (30a).

The inequality in (30b) follows using $T_{k+1} \leq \mathcal{T}_k$ since

$$\begin{aligned}
&W_{k+1} - W_k \\
&= \mathcal{T}_{k+1}(f(X_{k+1}) - f_{\inf} + r(X_{k+1}) - r_{\inf}) + \|c(X_{k+1})\|_2 \\
&\quad - (\mathcal{T}_k(f(X_k) - f_{\inf} + r(X_k) - r_{\inf}) + \|c(X_k)\|_2) \\
&\leq \mathcal{T}_k(f(X_{k+1}) - f_{\inf} + r(X_{k+1}) - r_{\inf}) + \|c(X_{k+1})\|_2 \\
&\quad - (\mathcal{T}_k(f(X_k) - f_{\inf} + r(X_k) - r_{\inf}) + \|c(X_k)\|_2) \\
&= \mathcal{T}_k(f(X_{k+1}) + r(X_{k+1})) + \|c(X_{k+1})\|_2 - (\mathcal{T}_k(f(X_k) + r(X_k)) + \|c(X_k)\|_2) \\
&= \phi(X_{k+1}, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k),
\end{aligned}$$

which completes the proof of the lemma. □ □

We now summarize key properties of the stochastic process.

Lemma 3.12. *Suppose that ϵ satisfies* (21)*. Moreover, define $\omega : \mathbb{R}^2_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $h : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ for all $(a, b) \in \mathbb{R}^2_{\geq 0}$ and $\alpha \in \mathbb{R}_{>0}$, respectively, by*

$$\begin{aligned}
\omega(a, b) &:= 2a + \tau_{-1} b \quad \textit{and} \\
h(\alpha) &:= \frac{\eta_1}{(\max\{\sqrt{2}, \kappa_T\})^2} \min\left\{\frac{1}{2(1+\kappa)^2}, \frac{1-2\eta}{2}\right\} \tau_{\min} \alpha \epsilon^2.
\end{aligned}$$

Then, the following hold.

(i) One has that

$$h(\bar{\alpha}) > \frac{\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f)}{p - \frac{1}{2}} \quad \textit{and} \quad p \in \left(\frac{1}{2} + \frac{\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f)}{h(\bar{\alpha})}, 1\right]. \tag{31}$$

(ii) Conditioned on $k < T_\epsilon$, the following hold.

(a) Conditioned on $I_k \Theta_k = 1$, one has

$$W_{k+1} \leq W_k - h(A_k) + \omega(\bar{\epsilon}_f, E_k + E_k^+) \leq W_k - h(A_k) + \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f).$$

(b) Conditioned on $A_k \leq \bar{\alpha}$ and $I_k = 1$, one has $\Theta_k = 1$.

(c) $W_{k+1} \leq W_k + \omega(\bar{\epsilon}_f, E_k + E_k^+)$.

(iii) For all $k \in \mathbb{N}$, it holds that $\mathbb{P}_{M_k}[I_k = 1 | \mathcal{F}_k] \geq p$.

Proof. Consider part (i). It follows along with (21) and $\theta \in (0, 1)$ that

$$\begin{aligned}
h(\bar{\alpha}) &= \frac{\eta_1}{(\max\{\sqrt{2}, \kappa_T\})^2} \min\left\{\frac{1}{2(1+\kappa)^2}, \frac{1-2\eta}{2}\right\} \tau_{\min} \bar{\alpha} \epsilon^2 \\
&> \frac{\eta_1}{2(\max\{\sqrt{2}, \kappa_T\})^2} \min\left\{\frac{1}{(1+\kappa)^2}, 1-2\eta\right\} \tau_{\min} \bar{\alpha} \\
&\quad \cdot (\max\{\sqrt{2}, \kappa_T\})^2 \max\left\{(1+\kappa)^2, \frac{1}{1-2\eta}\right\} \frac{(4+4\tau_{-1})\bar{\epsilon}_f}{\theta \tau_{\min} \bar{\alpha} \eta_1 (p - 1/2)} \\
&= \frac{1}{2\theta} \frac{(4+4\tau_{-1})\bar{\epsilon}_f}{p - 1/2} > \frac{(2+2\tau_{-1})\bar{\epsilon}_f}{p - 1/2} = \frac{\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f)}{p - 1/2},
\end{aligned}$$

as claimed. After rearrangement, this inequality along with the bounds in (17) together imply that p is on the stated interval in part (i).

Now consider part (ii) and condition on the event that $k < T_\epsilon$. Let us begin by proving part (ii)(a), for which we condition on the event that $I_k \Theta_k = 1$. Thus, iteration k is true, so the inequalities in Definition 3.3 hold. We consider two cases.

- Suppose that $\|G_k - \nabla f(X_k)\|_2 \leq \kappa\sqrt{\|S_k\|_2^2 + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}$. Thus, by Lemma 3.10, the conclusion of part (i) of that lemma holds. Combining this conclusion with the fact that iteration k is also successful and $\mathcal{T}_k \leq 1$,

$$\begin{aligned}
&\Phi_k^+ - \Phi_k - 2\bar{\epsilon}_f \\
\leq\ & -\eta_1\left(\mathcal{T}_k\frac{\|S_k\|_2^2}{2A_k} + (\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2)\right) \\
\leq\ & -\frac{\eta_1}{(1+\kappa)^2}\left(\mathcal{T}_k\frac{\|S_k^{\text{true}}\|_2^2}{2A_k} + (\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2)\right) \\
\leq\ & -\frac{\eta_1}{(1+\kappa)^2}\mathcal{T}_k\left(\frac{\|S_k^{\text{true}}\|_2^2}{2A_k} + (\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2)\right) \\
=\ & -\frac{\eta_1}{(1+\kappa)^2}\mathcal{T}_k A_k\left(\frac{\|S_k^{\text{true}}\|_2^2}{2A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}\right) \\
\leq\ & -\frac{\eta_1}{2(1+\kappa)^2}\mathcal{T}_k A_k\left(\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}\right).
\end{aligned}$$

- Suppose that $\|G_k - \nabla f(X_k)\|_2 \leq \epsilon_g$. Thus, by Lemma 3.10, the conclusion of part (ii) of that lemma holds. Combining this conclusion with the fact that iteration k is also successful and $\mathcal{T}_k \leq 1$,

$$\begin{aligned}
&\Phi_k^+ - \Phi_k - 2\bar{\epsilon}_f \\
\leq\ & -\eta_1\left(\mathcal{T}_k\frac{\|S_k\|_2^2}{2A_k} + (\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2)\right) \\
\leq\ & -\eta_1(1-2\eta)\left(\mathcal{T}_k\frac{\|S_k^{\text{true}}\|_2^2}{2A_k} + (\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2)\right) \\
\leq\ & -\eta_1(1-2\eta)\mathcal{T}_k\left(\frac{\|S_k^{\text{true}}\|_2^2}{2A_k} + (\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2)\right) \\
=\ & -\eta_1(1-2\eta)\mathcal{T}_k A_k\left(\frac{\|S_k^{\text{true}}\|_2^2}{2A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}\right) \\
\leq\ & -\eta_1\frac{1-2\eta}{2}\mathcal{T}_k A_k\left(\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}\right).
\end{aligned}$$

Combining these inequalities, Lemma 3.3, $A_k \leq 1$, and $\|S_k^{\text{true}}\|_2 \geq \|\min\{X_k, -Z_k^{\text{true}}\}\|_2$ (see [7, Lemma 4.6]) one finds with $\chi := \min\left\{\frac{1}{2(1+\kappa)^2}, \frac{1-2\eta}{2}\right\}$ that

$$\begin{aligned}
&\Phi_k^+ - \Phi_k - 2\bar{\epsilon}_f \\
\leq\ & -\eta_1\chi\mathcal{T}_k A_k\left(\frac{\|S_k^{\text{true}}\|_2^2}{A_k^2} + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}\right) \\
\leq\ & -\eta_1\chi\mathcal{T}_k A_k\Bigg(\frac{\|S_k^{\text{true}}\|_2^2}{2A_k^2} + \frac{\|\min\{X_k, -Z_k^{\text{true}}\}\|_2^2}{2A_k^2} \\
& \qquad + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}\Bigg) \\
\leq\ & -\eta_1\chi\tau_{\min} A_k\Bigg(\frac{\|S_k^{\text{true}}\|_2^2}{2A_k^2} + \frac{\|\min\{X_k, -Z_k^{\text{true}}\}\|_2^2}{2} \\
& \qquad + \frac{\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}{A_k}\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&\leq -\frac{\eta_1}{(\max\{\sqrt{2},\kappa_T\})^2}\chi\tau_{\min}A_k\epsilon^2 \\
&= -h(A_k).
\end{aligned}
$$

Hence, it follows with (30a), nonnegativity of E_k and E_k^+, and $\mathcal{T}_k \leq \tau_{-1}$ that

$$
\begin{aligned}
\phi(X_k^+,\mathcal{T}_k)-\phi(X_k,\mathcal{T}_k) &\leq \Phi_k^+ - \Phi_k + \mathcal{T}_k(E_k+E_k^+) \\
&\leq -h(A_k) + 2\bar{\epsilon}_f + \tau_{-1}(E_k+E_k^+) \\
&= -h(A_k) + \omega(\bar{\epsilon}_f, E_k+E_k^+).
\end{aligned}
$$

Combining this inequality, (30b), $\Theta_k = 1$, the fact that $\omega(\cdot,\cdot)$ is an increasing function in its second argument, and $E_k + E_k^+ \leq 2\bar{\epsilon}_f$ give

$$
\begin{aligned}
W_{k+1} - W_k \leq \phi(X_{k+1},\mathcal{T}_k) - \phi(X_k,\mathcal{T}_k) &= \phi(X_k^+,\mathcal{T}_k) - \phi(X_k,\mathcal{T}_k) \\
&\leq -h(A_k) + \omega(\bar{\epsilon}_f, E_k+E_k^+) \leq -h(A_k) + \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f),
\end{aligned}
$$

which completes the proof of part (ii)(a).

Now consider part (ii)(b), for which we condition on the event that $A_k \leq \bar{\alpha}$ and $I_k = 1$. Thus, the inequalities in Definition 3.3 hold. Let us consider two cases.

- Suppose that $\|G_k - \nabla f(X_k)\|_2 \leq \kappa\sqrt{\|S_k\|_2^2 + \|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2}$. Let us first observe from $\tau_{\min} \leq \mathcal{T}_k$ and $A_k \leq \bar{\alpha} \leq \omega_1^{-1}$ that

$$
\begin{aligned}
&1 - \left(\frac{\mathcal{T}_k L_g + L_J}{\mathcal{T}_k}\right)A_k - \left(\frac{\kappa^2}{2\eta_1} + 2\eta_1\right)A_k \\
&\geq 1 - \left(\frac{\tau_{\min} L_g + L_J}{\tau_{\min}} + \frac{\kappa^2}{2\eta_1} + 2\eta_1\right)A_k \\
&\geq 1 - (1-\eta_1)\left(\frac{\tau_{\min}(L_g+2\eta_1)+L_J}{\tau_{\min}} + \frac{\kappa^2}{2\eta_1}\right)\left(\frac{\tau_{\min}(L_g+2\eta_1)+L_J}{\tau_{\min}} + \frac{\kappa^2}{2\eta_1}\right)^{-1} \\
&= \eta_1.
\end{aligned}
$$

  Combining this inequality with Lemma 3.9(i) gives

$$
\begin{aligned}
&\phi(X_k+S_k,\mathcal{T}_k) - \phi(X_k,\mathcal{T}_k) \\
&\leq -\frac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + \frac{\mathcal{T}_k L_g + L_J}{2}\|S_k\|_2^2 + \left(\frac{\kappa^2}{4\eta_1} + \eta_1\right)A_k\mathcal{T}_k\frac{\|S_k\|_2^2}{A_k} \\
&\quad - \eta_1(\|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2) \\
&= -\left(1 - \left(\frac{\mathcal{T}_k L_g + L_J}{\mathcal{T}_k}\right)A_k - \left(\frac{\kappa^2}{2\eta_1} + 2\eta_1\right)A_k\right)\frac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 \\
&\quad - \eta_1(\|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2) \\
&\leq -\eta_1\frac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 - \eta_1(\|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2) \\
&= -\eta_1\left(\frac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + (\|c(X_k)\|_2 - \|c(X_k)+\nabla c(X_k)^T V_k\|_2)\right).
\end{aligned}
$$

- Suppose that $\|G_k - \nabla f(X_k)\|_2 \leq \epsilon_g$. From $\tau_{\min} \leq \mathcal{T}_k$ and $A_k \leq \bar{\alpha} \leq \omega_2^{-1}$,

$$
1 - \left(\frac{\eta^2}{2\eta_1(1-2\eta)^2} + 2\eta_1\right) - \frac{(\mathcal{T}_k L_g + L_J)A_k}{\mathcal{T}_k}
$$

$$
\begin{aligned}
&\geq 1 - \left(\frac{\eta^2}{2\eta_1(1-2\eta)^2} + 2\eta_1\right) - \frac{(\tau_{\min} L_g + L_J)\omega_2^{-1}}{\tau_{\min}} \\
&= 1 - \frac{\eta^2}{2\eta_1(1-2\eta)^2} - 2\eta_1 - \left(1 - \frac{\eta^2}{2\eta_1(1-2\eta)^2} - 3\eta_1\right) \\
&= \eta_1.
\end{aligned}
$$

Combining this inequality with Lemma 3.9(ii) gives

$$
\begin{aligned}
&\phi(X_k + S_k, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k) \\
\leq\ & -\frac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + \frac{\mathcal{T}_k L_g + L_J}{2}\|S_k\|_2^2 + \left(\frac{\eta^2}{4\eta_1(1-2\eta)^2} + \eta_1\right)\mathcal{T}_k \frac{\|S_k\|_2^2}{A_k} \\
& - \eta_1(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2) \\
=\ & -\left(1 - \left(\frac{\eta^2}{2\eta_1(1-2\eta)^2} + 2\eta_1\right) - \left(\frac{\mathcal{T}_k L_g + L_J}{\mathcal{T}_k}\right) A_k\right)\frac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 \\
& - \eta_1(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2) \\
\leq\ & -\eta_1 \frac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 - \eta_1(\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2) \\
=\ & -\eta_1\left(\frac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + (\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2)\right).
\end{aligned}
$$

It follows from the previous two cases, (30a), $E_k + E_k^+ \leq 2\bar{\epsilon}_f$, and $\mathcal{T}_k \leq 1$ that

$$
\begin{aligned}
\Phi_k^+ - \Phi_k &\leq \phi(X_k^+, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k) + \mathcal{T}_k(E_k + E_k^+) \leq \phi(X_k^+, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k) + 2\bar{\epsilon}_f \\
&\leq -\eta_1\left(\frac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + (\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2)\right) + 2\bar{\epsilon}_f,
\end{aligned}
$$

which implies that $\Theta_k = 1$, thus completing the proof of part (ii)(b).

Now consider part (ii)(c). Note that $W_{k+1} \leq W_k$ if the kth iteration is not successful since $\mathcal{T}_{k+1} \leq \mathcal{T}_k$ and $X_{k+1} = X_k$, in which case the inequality in part (ii)(c) holds trivially. Thus, for the remainder of the proof of this part, let us condition on the event that $\Theta_k = 1$. In this case, one finds that

$$
\Phi_k^+ - \Phi_k \leq -\eta_1\left(\frac{\mathcal{T}_k}{2A_k}\|S_k\|_2^2 + (\|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2)\right) + 2\bar{\epsilon}_f \leq 2\bar{\epsilon}_f.
$$

It follows from this inequality, $\Theta_k = 1$, (30b), (30a), and $\mathcal{T}_k \leq \tau_{-1}$ that

$$
\begin{aligned}
W_{k+1} - W_k &\leq \phi(X_k^+, \mathcal{T}_k) - \phi(X_k, \mathcal{T}_k) \leq \Phi_k^+ - \Phi_k + \mathcal{T}_k(E_k + E_k^+) \\
&\leq 2\bar{\epsilon}_f + \tau_{-1}(E_k + E_k^+) = \omega(\bar{\epsilon}_f, E_k + E_k^+),
\end{aligned}
$$

which completes the proof of part (ii)(c).

Finally, consider part (iii). Letting $1\{A\}$ denote the indicator function for the event A, let us define for convenience the random variables

$$
\begin{aligned}
I_{k,1} &:= 1\Bigg\{\|G_k - \nabla f(X_k)\| \\
&\qquad \leq \max\left\{\epsilon_g, \kappa\sqrt{\|S_k\|_2^2 + \|c(X_k)\|_2 - \|c(X_k) + \nabla c(X_k)^T V_k\|_2}\right\}\Bigg\} \\
\text{and}\quad I_{k,2} &:= 1\left\{E_k + E_k^+ \leq 2\bar{\epsilon}_f\right\}.
\end{aligned}
$$

It follows from these definitions and Assumption 3.2 that

$$
\mathbb{P}_{M_k}[I_k = 0 | \mathcal{F}_k] = \mathbb{P}_{M_k}[I_{k,1} = 0 \text{ or } I_{k,2} = 0 | \mathcal{F}_k]
$$

$$\leq \mathbb{P}_{M_k}[I_{k,1} = 0|\mathcal{F}_k] + \mathbb{P}_{M_k}[I_{k,2} = 0|\mathcal{F}_k]$$
$$\leq \delta + \mathbb{P}_{M_k}[I_{k,2} = 0|\mathcal{F}_k]. \tag{32}$$

We now bound the probability on the right-hand side of (32). It follows from (14), (15), $M_k = (\Xi_{k,g}, \Xi_{k,f}, \Xi^+_{k,f})$, linearity of the expectation operator, and the Cauchy-Schwarz inequality for expectations that

$$\mathbb{E}_{M_k}[E_k + E_k^+|\mathcal{F}_k] \leq 2\epsilon_f \quad \text{and}$$
$$\mathbb{E}_{M_k}\left[\exp\left\{\lambda\left(E_k + E_k^+ - \mathbb{E}_{M_k}[E_k + E_k^+|\mathcal{F}_k]\right)\right\}|\mathcal{F}_k\right] \leq \exp\left(\frac{\lambda^2(2\nu)^2}{2}\right)$$
$$\text{for all} \quad \lambda \in \left[0, \frac{1}{2b}\right].$$

Using a one-sided Bernstein inequality, it follows for all $t \geq 0$ that

$$\mathbb{P}_{M_k}[E_k + E_k^+ - \mathbb{E}_{M_k}[E_k + E_k^+|\mathcal{F}_k] \geq t|\mathcal{F}_k] \leq \exp\left(-\min\left\{\frac{t^2}{2(2\nu)^2}, \frac{t}{2 \cdot 2b}\right\}\right). \tag{33}$$

Note that

$$\mathbb{P}_{M_k}[I_{k,2} = 0|\mathcal{F}_k] = \mathbb{P}_{M_k}\Big[E_k + E_k^+ - \mathbb{E}_{M_k}[E_k + E_k^+|\mathcal{F}_k]$$
$$> 2\bar{\epsilon}_f - \mathbb{E}_{M_k}[E_k + E_k^+|\mathcal{F}_k]\Big|\mathcal{F}_k\Big]. \tag{34}$$

Next, let us define $\bar{t} := 2\bar{\epsilon}_f - \mathbb{E}_{M_k}[E_k + E_k^+ \,|\, \mathcal{F}_k]$, and note from linearity of the expectation operator, (14), (15), and Assumption 3.3 that $\bar{t} \geq 2u_f \geq 0$. Combining this inequality with (34) and (33), one obtains

$$\mathbb{P}_{M_k}[I_{k,2} = 0|\mathcal{F}_k] = \mathbb{P}_{M_k}[E_k + E_k^+ - \mathbb{E}_{M_k}[E_k + E_k^+|\mathcal{F}_k]] > \bar{t}|\mathcal{F}_k]$$
$$\leq \exp\left(-\min\left\{\frac{\bar{t}^2}{2(2\nu)^2}, \frac{\bar{t}}{2 \cdot 2b}\right\}\right)$$
$$\leq \exp\left(-\min\left\{\frac{4u_f^2}{2(2\nu)^2}, \frac{2u_f}{2 \cdot 2b}\right\}\right) = \exp\left(-\min\left\{\frac{u_f^2}{2\nu^2}, \frac{u_f}{2b}\right\}\right).$$

Combining this inequality with (32) gives

$$\mathbb{P}_{M_k}[I_k = 1|\mathcal{F}_k] = 1 - \mathbb{P}_{M_k}[I_k = 0|\mathcal{F}_k] \geq 1 - \delta - \exp\left(-\min\left\{\frac{u_f^2}{2\nu^2}, \frac{u_f}{2b}\right\}\right) = p,$$

which completes the proof. □ □

Our next result gives a concentration bound related to the sum of $\{I_k\}$.

Lemma 3.13. *Let $\hat{p} \in [0, p)$. Then, it holds that*

$$\mathbb{P}\left[\sum_{k=0}^{t-1} I_k \leq \hat{p}t\right] \leq \exp\left(-\frac{(p-\hat{p})^2}{2p^2}t\right) \quad \textit{for all } t \in \mathbb{N} \textit{ with } t \geq 1.$$

Proof. Consider arbitrary $\hat{p} \in [0, p)$. For all $t \in \mathbb{N}$ with $t \geq 1$ define the $\mathcal{F}_t$-measurable random variable $\overline{Y}_t := pt - \sum_{k=0}^{t-1} I_k$ and $\overline{Y}_0 := 0$. It follows that

$$\overline{Y}_{t+1} = p(t+1) - \sum_{k=0}^{t} I_k = pt + p - I_t - \sum_{k=0}^{t-1} I_k = \overline{Y}_t + p - I_t. \tag{35}$$

Combining (35), linearity of expectation, and Lemma 3.12(iii) gives

$$\begin{aligned}\mathbb{E}_{M_t}[\overline{Y}_{t+1}|\mathcal{F}_t] &= \mathbb{E}_{M_t}[\overline{Y}_t + p - I_t|\mathcal{F}_t] = \overline{Y}_t + p - \mathbb{E}_{M_t}[I_t|\mathcal{F}_t] \\ &= \overline{Y}_t + p - \mathbb{P}_{M_t}[I_t = 1|\mathcal{F}_t] \leq \overline{Y}_t + p - p = \overline{Y}_t,\end{aligned}$$

from which it follows that $\{\overline{Y}_t\}$ is a supermartingale. Next, it follows from (35), the fact that $I_k \in \{0, 1\}$ for all $k \in \mathbb{N}$, and $p > 1/2$ (see (17)) that

$$|\overline{Y}_t - \overline{Y}_{t-1}| = |p - I_{t-1}| \leq \max\{p, 1-p\} = p.$$

Thus, each increment of the supermartingale $\{\overline{Y}_t\}$ is bounded in absolute value by p. Hence, the Azuma–Hoeffding inequality [1] gives

$$\begin{aligned}\mathbb{P}\left[\sum_{k=0}^{t-1} I_k \leq \hat{p}t\right] &= \mathbb{P}\left[pt - \sum_{k=0}^{t-1} I_k \geq (p-\hat{p})t\right] = \mathbb{P}[\overline{Y}_t \geq (p-\hat{p})t] \\ &\leq \exp\left(-\frac{(p-\hat{p})^2 t^2}{2tp^2}\right) = \exp\left(-\frac{(p-\hat{p})^2}{2p^2}t\right),\end{aligned}$$

which completes the proof. □ □

Next, we show that the number of successful iterations with a "large" proximal parameter (Definition 3.7) is at least as large as the number of unsuccessful iterations with a "large" proximal parameter minus a constant. The result and its proof are similar to [14, Lemma 3.2] with a small modification to handle the upper bound enforced on the proximal parameter in Algorithm 1, which was not used in [14].

Lemma 3.14. *With $\bar{i}$ from Definition 3.5, it holds for all $t \in \mathbb{N} \setminus \{0\}$ that*

$$\sum_{k=0}^{t-1} L_k\Theta_k \geq \sum_{k=0}^{t-1} 1\{A_k < \alpha_0\}L_k\Theta_k \geq \sum_{k=0}^{t-1} L_k(1-\Theta_k) - (\bar{i}+1).$$

Proof. The first inequality is trivial. Next, since a large proximal parameter in iteration k has A_k and A_{k+1} bounded below by $\bar{\alpha}$, the number of large proximal parameters for which A_k gets decreased is bounded by the number of large proximal parameters with $A_k < \alpha_0$ for which A_k gets increased, plus $\bar{i}+1$, which is the number of unsuccessful steps it takes to decrease the proximal parameter from α_0 to $\bar{\alpha}$. This establishes the second inequality above. □ □

The next result is the same as [14, Corollary 3.3].

Lemma 3.15. *For all $t \in \mathbb{N} \setminus \{0\}$, it holds that $\sum_{k=0}^{t-1} L_k\Theta_k \geq \frac{1}{2}\left(\sum_{k=0}^{t-1} L_k - (\bar{i}+1)\right)$.*

Proof. Use the identity $L_k(1-\Theta_k) = L_k - L_k\Theta_k$ in Lemma 3.14 and rearrange. □ □

The next result is the same as [14, Lemma 3.4], with additional details of the proof provided here for enhanced clarity.

Lemma 3.16. *Conditioned on $t < T_\epsilon$, we have $\sum_{k=0}^{t-1}(1-L_k)I_k \leq \sum_{k=0}^{t-1}(1-L_k)(1-I_k)$.*

Proof. It follows from Lemma 3.12(ii)(b) and the fact that $A_k \leq \bar{\alpha}$ when $L_k = 0$ that

$$\sum_{k=0}^{t-1}(1-L_k)I_k \leq \sum_{k=0}^{t-1}(1-L_k)\Theta_k \quad \text{and} \quad \sum_{k=0}^{t-1}(1-L_k)(1-\Theta_k) \leq \sum_{k=0}^{t-1}(1-L_k)(1-I_k). \tag{36}$$

Also, since the number of small proximal parameters where A_k is increased is bounded by the number of small proximal parameters where A_k is decreased,

$$\sum_{k=0}^{t-1}(1-L_k)\Theta_k \leq \sum_{k=0}^{t-1}(1-L_k)(1-\Theta_k). \tag{37}$$

The claimed result follows by combining (36) and (37). □ □

We can now prove that the probability that a certain event occurs is zero.

Lemma 3.17. *Let $\hat{p} \in (\frac{1}{2}, 1]$. Then, for all $t \in \mathbb{N} \setminus \{0\}$, it holds that*

$$\mathbb{P}\left[T_\epsilon > t,\ \sum_{k=0}^{t-1} I_k \geq \hat{p}t,\ \textit{and}\ \sum_{k=0}^{t-1} L_k\Theta_k I_k < \left(\hat{p}-\frac{1}{2}\right)t - \frac{(\bar{i}+1)}{2}\right] = 0.$$

Proof. The proof follows exactly as the proof for [14, Lemma 3.5], which uses [14, Corollary 3.3]) (equivalent to our Lemma 3.15), [14, Lemma 3.4] (equivalent to our Lemma 3.16), and the notation U_k (analogous to our notation L_k). □ □

We now prove a worst-case iteration complexity result.

Theorem 3.18. *Suppose that Assumptions 1.1–3.5 hold, let T_ϵ be defined as in Definition 3.5. Then, the set of $(s, \hat{p}, t) \in \mathbb{R} \times \mathbb{R} \times \mathbb{N}$ satisfying*

$$s \in \left(0, \left(p-\tfrac{1}{2}\right)h(\bar{\alpha}) - \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f)\right), \tag{38a}$$

$$\hat{p} \in \left(\frac{1}{2} + \frac{\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f)+s}{h(\bar{\alpha})}, p\right), \quad \textit{and} \tag{38b}$$

$$t \geq \hat{t} := \frac{\overline{W}_0/h(\bar{\alpha}) + (\bar{i}+1)/2}{\hat{p} - \left(\frac{1}{2} + \frac{\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f)+s}{h(\bar{\alpha})}\right)} \tag{38c}$$

is nonempty, where $\overline{W}_0 := \tau_{-1}\big(f(x_0) - f_{\inf} + r(x_0) - r_{\inf}\big) + \|c(x_0)\|_2 \geq W_0$. Moreover, for all $(s, \hat{p}, t)$ satisfying (38)*, it holds that*

$$\mathbb{P}[T_\epsilon \leq t] \geq 1 - \exp\left(-\frac{(p-\hat{p})^2}{2p^2}t\right) - \exp\left(-\min\left\{\frac{s^2 t}{2(2\tau_{-1}\nu)^2}, \frac{st}{2(2\tau_{-1}b)}\right\}\right), \tag{39}$$

where the right-hand side of (39) *is positive for all t sufficiently large. Thus, Lemma 3.4 gives for all such t that, with probability at least the right-hand side of* (39)*, Algorithm 1 computes an iteration $\bar{k} \leq t$ so that $x_{\bar{k}}$ is an ϵ-KKT point for problem* (1)*.*

Proof. The interval in (38a) is nonempty as a consequence of (31). Next, let s be any real number satisfying (38a), which implies that $0 < s < \left(p-\frac{1}{2}\right)h(\bar{\alpha}) - \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f)$. After rearranging the second inequality, one obtains $\frac{1}{2} + \frac{\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f)+s}{h(\bar{\alpha})} < p$ so that the interval in (38b) is nonempty. Finally, let $\hat{p}$ be any real number satisfying (38b). The denominator in (38c) is positive and finite since $\hat{p}$ satisfies (38b), and the numerator is nonnegative and finite by the definitions of $\overline{W}_0$, $h(\cdot)$, $\bar{\alpha}$, and $\bar{i}$. Therefore, the right-hand side of (38c) is nonnegative and finite, meaning that all t sufficiently large satisfy (38c), which completes the proof of the first claim.

Next, let $(s, \hat{p}, t)$ satisfy (38). By the law of total probability,

$$\mathbb{P}[T_\epsilon > t] = \mathbb{P}\left[\underbrace{T_\epsilon > t \text{ and } \frac{1}{t}\sum_{k=0}^{t-1}\omega(\bar{\epsilon}_f, E_k + E_k^+) > \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) + s}_{\mathcal{A}}\right]$$

$$+\mathbb{P}\left[\underbrace{T_\epsilon > t \text{ and } \frac{1}{t}\sum_{k=0}^{t-1}\omega(\bar{\epsilon}_f, E_k + E_k^+) \le \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) + s}_{\mathcal{B}}\right].$$

Let us now bound $\mathbb{P}[\mathcal{A}]$ and $\mathbb{P}[\mathcal{B}]$ in order to prove (39).

First, consider $\mathbb{P}[\mathcal{A}]$. For any $k \in \mathbb{N}$, Assumption 3.1, Assumption 3.3 (specifically, $u_f \ge 0$), (15), $M_k = (\Xi_{k,g}, \Xi_{k,f}, \Xi_{k,f}^+)$, and linearity of expectation show that

$$\begin{aligned} &\mathbb{E}_{M_k}[\tau_{-1}(E_k + E_k^+)|\mathcal{F}_k] \le 2\tau_{-1}\epsilon_f \le 2\tau_{-1}\bar{\epsilon}_f \quad \text{and} \\ &\mathbb{E}_{M_k}\left[\exp\left\{\lambda\big(\tau_{-1}(E_k + E_k^+) - \mathbb{E}_{M_k}[\tau_{-1}(E_k + E_k^+)|\mathcal{F}_k]\big)\right\}|\mathcal{F}_k\right] \\ \le\ &\exp\left(\frac{\lambda^2(2\tau_{-1}\nu)^2}{2}\right) \quad \text{for all} \quad \lambda \in \left[0, \frac{1}{2\tau_{-1}b}\right]. \end{aligned} \tag{40}$$

Next, (40) and letting $\mathcal{E}_k := \tau_{-1}(E_k + E_k^+) - \mathbb{E}_{M_k}[\tau_{-1}(E_k + E_k^+)\,|\,\mathcal{F}_k]$ gives

$$\begin{aligned} \mathbb{P}[\mathcal{A}] &\le \mathbb{P}\left[\frac{1}{t}\sum_{k=0}^{t-1}\omega(\bar{\epsilon}_f, E_k + E_k^+) > \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) + s\right] \\ &= \mathbb{P}\left[\frac{1}{t}\sum_{k=0}^{t-1}\big(2\bar{\epsilon}_f + \tau_{-1}(E_k + E_k^+)\big) > 2\bar{\epsilon}_f + 2\tau_{-1}\bar{\epsilon}_f + s\right] \\ &= \mathbb{P}\left[\frac{1}{t}\sum_{k=0}^{t-1}(\tau_{-1}(E_k + E_k^+) - 2\tau_{-1}\bar{\epsilon}_f) > s\right] \\ &\le \mathbb{P}\left[\frac{1}{t}\sum_{k=0}^{t-1}(\tau_{-1}(E_k + E_k^+) - \mathbb{E}_{M_k}[\tau_{-1}(E_k + E_k^+)|\mathcal{F}_k]) > s\right] \\ &= \mathbb{P}\left[\frac{1}{t}\sum_{k=0}^{t-1}\mathcal{E}_k > s\right] \le \mathbb{P}\left[\sum_{k=0}^{t-1}\mathcal{E}_k \ge st\right]. \end{aligned} \tag{41}$$

Also, it follows from the tower property and (40) that

$$\begin{aligned} \mathbb{E}\left[\exp\left(\lambda\sum_{k=0}^{t-1}\mathcal{E}_k\right)\right] &= \mathbb{E}\left[\mathbb{E}_{M_{t-1}}\left[\exp\left(\lambda\sum_{k=0}^{t-2}\mathcal{E}_k\right)\exp\left(\lambda\mathcal{E}_{t-1}\right)\Big|\mathcal{F}_{t-1}\right]\right] \\ &\le \exp\left(\frac{\lambda^2(2\tau_{-1}\nu)^2}{2}\right)\mathbb{E}\left[\exp\left(\lambda\sum_{k=0}^{t-2}\mathcal{E}_k\right)\right] \quad \text{for all } \lambda \in \left[0, \frac{1}{2\tau_{-1}b}\right]. \end{aligned} \tag{42}$$

We can now use the Markov inequality and an iterative application of (42) to get

$$\begin{aligned} \mathbb{P}\left[\sum_{k=0}^{t-1}\mathcal{E}_k \ge st\right] &\le \mathbb{P}\left[\exp\left(\lambda\sum_{k=0}^{t-1}\mathcal{E}_k\right) \ge \exp(\lambda st)\right] \le \exp(-\lambda st)\mathbb{E}\left[\exp\left(\lambda\sum_{k=0}^{t-1}\mathcal{E}_k\right)\right] \\ &\le \exp(-\lambda st)\exp\left(\frac{t\lambda^2(2\tau_{-1}\nu)^2}{2}\right) \\ &= \exp\left(-\lambda st + \frac{t(2\tau_{-1}\nu)^2\lambda^2}{2}\right) \quad \text{for all } \lambda \in \left[0, \frac{1}{2\tau_{-1}b}\right]. \end{aligned}$$

We now combine this result with (41) to obtain

$$\mathbb{P}[\mathcal{A}] \le \exp\left(-\lambda st + \frac{t(2\tau_{-1}\nu)^2\lambda^2}{2}\right) =: h_{\mathcal{A}}(\lambda) \quad \text{for all } \lambda \in \left[0, \frac{1}{2\tau_{-1}b}\right]. \tag{43}$$

We now minimize $h_{\mathcal{A}}(\cdot)$ over all λ in (43). The *unconstrained* minimizer of $h_{\mathcal{A}}$ is

$$\lambda_u^* := \frac{s}{(2\tau_{-1}\nu)^2}.$$

Case 1. If $\lambda_u^* < \frac{1}{2\tau_{-1}b}$, i.e., if $s < \frac{(2\tau_{-1}\nu)^2}{2\tau_{-1}b}$, then the minimizer of $h_{\mathcal{A}}$ is $\lambda^* = \lambda_u^*$ with $h_{\mathcal{A}}(\lambda^*) = \exp\left(-\frac{s^2 t}{2(2\tau_{-1}\nu)^2}\right)$.

Case 2. If $\lambda_u^* \geq \frac{1}{2\tau_{-1}b}$, i.e., if $s \geq \frac{(2\tau_{-1}\nu)^2}{2\tau_{-1}b}$, then the *constrained* minimizer of $h_{\mathcal{A}}$ is $\lambda^* = \frac{1}{2\tau_{-1}b}$ with $h_{\mathcal{A}}(\lambda^*) = \exp\left(-\frac{st}{2\tau_{-1}b} + \frac{t(\tau_{-1}\nu)^2}{2(\tau_{-1}b)^2}\right)$. Also note that since $s \geq \frac{(2\tau_{-1}\nu)^2}{2\tau_{-1}b} = \frac{2(\tau_{-1}\nu)^2}{\tau_{-1}b}$, one has $\frac{(\tau_{-1}\nu)^2}{\tau_{-1}b} \leq \frac{s}{2}$. Combining these two results shows

$$\begin{aligned} h_{\mathcal{A}}(\lambda^*) &= \exp\left(-\frac{st}{2\tau_{-1}b} + \frac{t(\tau_{-1}\nu)^2}{2(\tau_{-1}b)^2}\right) = \exp\left(-\frac{st}{2\tau_{-1}b} + \frac{t}{2\tau_{-1}b}\cdot\frac{(\tau_{-1}\nu)^2}{\tau_{-1}b}\right) \\ &\leq \exp\left(-\frac{st}{2\tau_{-1}b} + \frac{st}{4\tau_{-1}b}\right) = \exp\left(-\frac{st}{4\tau_{-1}b}\right). \end{aligned}$$

Combining the above two cases gives

$$\begin{aligned} \mathbb{P}[\mathcal{A}] &\leq \max\left\{\exp\left(-\frac{s^2 t}{2(2\tau_{-1}\nu)^2}\right), \exp\left(-\frac{st}{4\tau_{-1}b}\right)\right\} \\ &= \exp\left(-\min\left\{\frac{s^2 t}{2(2\tau_{-1}\nu)^2}, \frac{st}{2(2\tau_{-1}b)}\right\}\right), \end{aligned} \tag{44}$$

which is our desired bound on $\mathbb{P}[\mathcal{A}]$.

To bound $\mathbb{P}[\mathcal{B}]$, first, the law of total probability gives

$$\begin{aligned} &\mathbb{P}[\mathcal{B}] \\ &= \mathbb{P}\left[\underbrace{T_\epsilon > t,\ \frac{1}{t}\sum_{k=0}^{t-1}\omega(\bar{\epsilon}_f, E_k + E_k^+) \leq \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) + s,\ \text{and}\ \sum_{k=0}^{t-1}\Theta_k I_k L_k < \left(\hat{p} - \frac{1}{2}\right)t - \frac{(\bar{i}+1)}{2}}_{\mathcal{B}_1}\right] \\ &\quad + \mathbb{P}\left[\underbrace{T_\epsilon > t,\ \frac{1}{t}\sum_{k=0}^{t-1}\omega(\bar{\epsilon}_f, E_k + E_k^+) \leq \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) + s,\ \text{and}\ \sum_{k=0}^{t-1}\Theta_k I_k L_k \geq \left(\hat{p} - \frac{1}{2}\right)t - \frac{(\bar{i}+1)}{2}}_{\mathcal{B}_2}\right]. \end{aligned}$$

It follows from the definition of h and the fact that $L_k = 1$ implies $A_k \geq \bar{\alpha}$ that $L_k = 1$ also implies $h(A_k) \geq h(\bar{\alpha})$. One can combine this fact with Lemma 3.12(ii)(a) and Lemma 3.12(ii)(c) to get, for all $k < T_\epsilon$, that

$$\begin{aligned} W_{k+1} &\leq W_k - h(\bar{\alpha}) + \omega(\bar{\epsilon}_f, E_k + E_k^+) \\ &\leq W_k - h(\bar{\alpha}) + \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) \ \text{ if } L_k I_k \Theta_k = 1, \text{ and} \\ W_{k+1} &\leq W_k + \omega(\bar{\epsilon}_f, E_k + E_k^+) \ \text{ if } L_k I_k \Theta_k = 0. \end{aligned} \tag{45}$$

One can now prove that $\mathbb{P}[\mathcal{B}_2] = 0$ by showing that

$$T_\epsilon > t \ \text{ and } \ \frac{1}{t}\sum_{k=0}^{t-1}\omega(\bar{\epsilon}_f, E_k + E_k^+) \leq \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) + s \tag{46}$$

together imply that

$$\sum_{k=0}^{t-1} \Theta_k I_k L_k < \left(\hat{p} - \frac{1}{2}\right) t - \frac{(\bar{i}+1)}{2}.$$

For a proof by contradiction, suppose that (46) holds and $\sum_{k=0}^{t-1} \Theta_k I_k L_k \geq \left(\hat{p} - \frac{1}{2}\right) t - \frac{(\bar{i}+1)}{2}$. This inequality, (45), $h(\bar{\alpha}) > 0$, and the second inequality in (46) show that

$$\begin{aligned}
W_t &= W_0 + \sum_{k=0}^{t-1} (W_{k+1} - W_k) \\
&= W_0 + \sum_{k=0}^{t-1} (W_{k+1} - W_k) 1\{L_k I_k \Theta_k = 1\} + \sum_{k=0}^{t-1} (W_{k+1} - W_k) 1\{L_k I_k \Theta_k = 0\} \\
&\leq W_0 + \sum_{k=0}^{t-1} \left(\omega(\bar{\epsilon}_f, E_k + E_k^+) - h(\bar{\alpha})\right) 1\{L_k I_k \Theta_k = 1\} \\
&\quad + \sum_{k=0}^{t-1} \omega(\bar{\epsilon}_f, E_k + E_k^+) 1\{L_k I_k \Theta_k = 0\} \\
&= W_0 - \sum_{k=0}^{t-1} 1\{L_k I_k \Theta_k = 1\} h(\bar{\alpha}) + \sum_{k=0}^{t-1} \omega(\bar{\epsilon}_f, E_k + E_k^+) \\
&\leq W_0 - \left[(\hat{p} - \frac{1}{2})t - \tfrac{\bar{i}+1}{2}\right] h(\bar{\alpha}) + t\left(\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) + s\right) \\
&= W_0 - \left(\left(\hat{p} - \frac{1}{2}\right) h(\bar{\alpha}) - \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) - s\right) t + \frac{(\bar{i}+1)}{2} h(\bar{\alpha}) \leq 0,
\end{aligned}$$

where the last inequality follows from (38c). This inequality and the definition of W_t gives $W_t = 0$, which means that X_t is a local minimizer to problem (1); in particular, $c(x_t) = 0$ and $X_t \in \Omega$. It now follows from (5) that $\Delta_t = 0$ so that $V_t = V_t^c = 0$ (see (6)). Combining this fact with Assumption 3.4 shows that $[\nabla c(X_t), I_{\mathcal{A}_t}^T]^T$ has full row rank, where $\mathcal{A}_t = \{i \in [n] : [X_t]_i = 0\}$. Thus, the LICQ holds for problem (1) at X_t. Since the LICQ holds at the minimizer X_t, one concludes that $U_t^{\text{true}} = 0$, which in turn implies $S_t^{\text{true}} = V_t + U_t^{\text{true}} = 0$. Combining this fact with the complementary slackness condition (see (8c)) for subproblem (7) with k replaced by t and g_k replaced by the true gradient $\nabla f(X_t)$, one finds that Z_t^{true} satisfies $\min\{X_t + V_t + U_t^{\text{true}}, -Z_t^{\text{true}}\} = \min\{X_t, -Z_t^{\text{true}}\} = 0$. To summarize, one has $S_t^{\text{true}} = 0$, $c(X_t) = 0$, $V_t = 0$, and $\min\{X_t, -Z_t^{\text{true}}\} = 0$, which means that $T_\epsilon \leq t$, which contradicts the first inequality in (46). Therefore, $\mathbb{P}[\mathcal{B}_2] = 0$.

Next, to bound $\mathbb{P}[\mathcal{B}_1]$, Lemma 3.13 and Lemma 3.17 give

$$\begin{aligned}
&\mathbb{P}[\mathcal{B}_1] \\
&\leq \mathbb{P}\left[T_\epsilon > t \ \text{ and } \ \sum_{k=0}^{t-1} \Theta_k I_k L_k < \left(\hat{p} - \frac{1}{2}\right) t - \frac{(\bar{i}+1)}{2}\right] \\
&= \mathbb{P}\left[T_\epsilon > t, \ \sum_{k=0}^{t-1} \Theta_k I_k L_k < \left(\hat{p} - \frac{1}{2}\right) t - \frac{(\bar{i}+1)}{2}, \ \text{ and } \ \sum_{k=0}^{t-1} I_k < \hat{p} t\right] \\
&\quad + \mathbb{P}\left[T_\epsilon > t, \ \sum_{k=0}^{t-1} \Theta_k I_k L_k < \left(\hat{p} - \frac{1}{2}\right) t - \frac{(\bar{i}+1)}{2}, \ \text{ and } \ \sum_{k=0}^{t-1} I_k \geq \hat{p} t\right] \\
&\leq \mathbb{P}\left[\sum_{k=0}^{t-1} I_k < \hat{p} t\right] + \mathbb{P}\left[T_\epsilon > t, \ \sum_{k=0}^{t-1} \Theta_k I_k L_k < \left(\hat{p} - \frac{1}{2}\right) t - \frac{(\bar{i}+1)}{2}, \ \text{ and } \ \sum_{k=0}^{t-1} I_k \geq \hat{p} t\right]
\end{aligned}$$

$$\leq \exp\left(-\frac{(p-\hat{p})^2}{2p^2}t\right) + 0 = \exp\left(-\frac{(p-\hat{p})^2}{2p^2}t\right). \tag{47}$$

Finally, one can use $\mathbb{P}[\mathcal{B}_2] = 0$ to obtain

$$\mathbb{P}[T_\epsilon \leq t] = 1 - \mathbb{P}[T_\epsilon > t] = 1 - \mathbb{P}[\mathcal{A}] - \mathbb{P}[\mathcal{B}] = 1 - \mathbb{P}[\mathcal{A}] - \mathbb{P}[\mathcal{B}_1].$$

The result (39) follows from this equation, (44), and (47). □ □

The parameter s measures how much "above expectation" the function evaluation error is allowed to be. From Lemma 3.12(ii)(c), we have $W_{k+1} - W_k \leq \omega(\bar{\epsilon}_f, E_k + E_k^+) = 2\bar{\epsilon}_f + \tau_{-1}(E_k + E_k^+)$. The parameter s serves as a bound on how far $\tau_{-1}(E_k + E_k^+)$ exceeds its mean bound $2\tau_{-1}\bar{\epsilon}_f$; a Bernstein-type inequality is then used to control the probability that the cumulative excess exceeds its mean bound over t iterations, yielding the second exponential in (39). Thus, s trades off two effects: a smaller s slows the decay of $\exp(-\min\{s^2t/(2(2\tau_{-1}\nu)^2), st/(2(2\tau_{-1}b))\})$ but enlarges the admissible range for $\hat{p}$ and lowers the threshold on t; and a larger s sharpens the concentration bound but shrinks the range for $\hat{p}$ (vanishing as $s \to (p - \frac{1}{2})h(\bar{\alpha}) - \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f)$) and raises the threshold on t. If $s = 0$ and we further assume, for all $k \in \mathbb{N}$, that $E_k + E_k^+ \leq 2\epsilon_f$ then the Bernstein-type term in (39) is no longer needed, and the concentration argument invoked on $E_k + E_k^+$ is also not needed. Therefore, in this case, the bound in (39) simplifies to

$$\mathbb{P}(T_\epsilon \leq t) \geq 1 - \exp\left(-\frac{(p-\hat{p})^2}{2p^2}t\right),$$

which is a high-probability iteration complexity in the bounded-noise setting.

From Theorem 3.18, we can derive an almost sure termination result and an expected iteration complexity result, both of which are stated in the next result.

Corollary 3.19. *Let the suppositions in Theorem 3.18 hold and suppose that ϵ satisfies* (21).

(i) An ϵ-KKT point is computed in a finite number of iterations with probability 1.

(ii) The expected number of iterations to reach an ϵ-KKT point is $O(\epsilon^{-2})$.

Proof. We first prove part (i). Fix any s and $\hat{p}$ satisfying (38a) and (38b), respectively, and let $\hat{t}$ be as defined in (38c). With $A_t := \{T_\epsilon > t\}$, Theorem 3.18 implies

$$\begin{aligned}
&\sum_{t=\lceil \hat{t} \rceil}^{\infty} \mathbb{P}(A_t) \\
&\leq \sum_{t=\lceil \hat{t} \rceil}^{\infty} \left[\exp\left(-\frac{(p-\hat{p})^2}{2p^2}\, t\right) + \exp\left(-\min\left\{\frac{s^2 t}{2(2\tau_{-1}\nu)^2}, \frac{st}{2(2\tau_{-1}b)}\right\}\right)\right] \\
&= \sum_{t=\lceil \hat{t} \rceil}^{\infty} \left\{\left[\underbrace{\exp\left(-\frac{(p-\hat{p})^2}{2p^2}\right)}_{<1}\right]^t + \left[\underbrace{\exp\left(-\min\left\{\frac{s^2}{2(2\tau_{-1}\nu)^2}, \frac{s}{2(2\tau_{-1}b)}\right\}\right)}_{<1}\right]^t\right\} \\
&< \infty,
\end{aligned}$$

which uses $\hat{p} < p$ and $s > 0$. It follows from this and the Borel-Cantelli lemma that $\mathbb{P}[\limsup_{t\to\infty} A_t] = 0$. Thus, $\mathbb{P}[T_\epsilon < \infty] = 1$, meaning that an ϵ-KKT point is found in a finite number of iterations with probability 1.

Next, consider part (ii). Let us define the following:

$$\rho := \tfrac{\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f)}{h(\bar{\alpha})}, \qquad \sigma := (p - \tfrac{1}{2}) - \rho, \tag{48}$$

$$M := \max\{\sqrt{2}, \kappa_T\}, \qquad \chi := \min\left\{\tfrac{1}{2(1+\kappa)^2}, \tfrac{1-2\eta}{2}\right\}, \tag{49}$$

and note that it follows from (49) that $h(\bar{\alpha})$ from Lemma 3.12 may be written as

$$h(\bar{\alpha}) \equiv \eta_1 M^{-2} \chi \tau_{\min} \bar{\alpha} \epsilon^2. \tag{50}$$

It follows from (50), (21), and $\max\left\{(1+\kappa)^2, \frac{1}{1-2\eta}\right\} = \frac{1}{2\chi}$ that

$$\frac{M^2 h(\bar{\alpha})}{\eta_1 \chi \tau_{\min} \bar{\alpha}} = \epsilon^2 > M^2 \frac{1}{2\chi} \frac{(4+4\tau_{-1})\bar{\epsilon}_f}{\theta \tau_{\min} \bar{\alpha} \eta_1 (p-1/2)} = M^2 \frac{1}{\chi} \frac{(2+2\tau_{-1})\bar{\epsilon}_f}{\theta \tau_{\min} \bar{\alpha} \eta_1 (p-1/2)},$$

which after simplification and using $\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) = (2+2\tau_{-1})\bar{\epsilon}_f$, one obtains

$$\rho = \frac{\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f)}{h(\bar{\alpha})} < \theta(p - \tfrac{1}{2}) \implies \sigma > (1-\theta)(p-\tfrac{1}{2}) > 0. \tag{51}$$

Next, let us choose for s, which is required to satisfy (38a), the specific value

$$s = \frac{\sigma h(\bar{\alpha})}{3} = \frac{(p - \frac{1}{2} - \rho) h(\bar{\alpha})}{3} = \frac{(p-\frac{1}{2}) h(\bar{\alpha}) - \omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f)}{3}, \tag{52}$$

which satisfies the requirement in (38a). With this choice of s, let us next choose for $\hat{p}$, which is required to satisfy (38b), the specific value

$$\hat{p} = \frac{1}{2} + \rho + \frac{2\sigma}{3}.$$

Note that the lower bound in (38b), with s chosen as above, satisfies

$$\frac{1}{2} + \frac{\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) + s}{h(\bar{\alpha})} = \frac{1}{2} + \frac{\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) + \sigma h(\bar{\alpha})/3}{h(\bar{\alpha})} = \frac{1}{2} + \rho + \frac{\sigma}{3} < \hat{p},$$

meaning that $\hat{p}$ as chosen above satisfies the lower bound in (38b). On the other hand, one has from the definition of σ, (51), and $\theta \in (0,1)$ that

$$\hat{p} = \frac{1}{2} + \rho + \frac{2\sigma}{3} = \frac{1}{2} + \frac{2}{3}\left(p - \frac{1}{2}\right) + \frac{1}{3}\rho < \frac{1}{2} + \left(\frac{2}{3} + \frac{1}{3}\theta\right)\left(p - \frac{1}{2}\right) < \frac{1}{2} + \left(p - \frac{1}{2}\right) = p$$

meaning that $\hat{p}$ also satisfies the upper bound in (38b). With these specific choices for s and $\hat{p}$, one can now compute the denominator of $\hat{t}$ in (38c) as being equal to

$$\begin{aligned} \Delta := \hat{p} - \left(\frac{1}{2} + \tfrac{\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) + s}{h(\bar{\alpha})}\right) &= \frac{1}{2} + \rho + \frac{2\sigma}{3} - \left(\frac{1}{2} + \tfrac{\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f) + \frac{\sigma h(\bar{\alpha})}{3}}{h(\bar{\alpha})}\right) \\ &= \frac{1}{2} + \rho + \frac{2\sigma}{3} - \left(\frac{1}{2} + \rho + \frac{\sigma}{3}\right) = \frac{\sigma}{3}, \end{aligned} \tag{53}$$

and can similarly calculate, for the choice of $\hat{p}$ above, that

$$p - \hat{p} = p - \frac{1}{2} - \rho - \frac{2\sigma}{3} = \sigma - \frac{2\sigma}{3} = \frac{\sigma}{3}. \tag{54}$$

One can now also see from (38c), (53), (50), (51), and (49) that

$$\begin{aligned} \hat{t} &= \frac{\overline{W}_0/h(\bar{\alpha}) + (\bar{i}+1)/2}{\Delta} = \frac{3}{\sigma}\left(\frac{\overline{W}_0}{\eta_1 M^{-2} \chi \tau_{\min} \bar{\alpha} \epsilon^2} + \frac{\bar{i}+1}{2}\right) \\ &< \frac{3}{(1-\theta)(p-\frac{1}{2})}\left(\frac{M^2 \overline{W}_0}{\eta_1 \chi \tau_{\min} \bar{\alpha} \epsilon^2} + \frac{\bar{i}+1}{2}\right) \\ &= \frac{3}{(1-\theta)(p-\frac{1}{2})}\left(\frac{\max\{2, \kappa_T^2\}\overline{W}_0}{\eta_1 \min\left\{\frac{1}{2(1+\kappa)^2}, \frac{1-2\eta}{2}\right\} \tau_{\min} \bar{\alpha} \epsilon^2} + \frac{\bar{i}+1}{2}\right) = O(\epsilon^{-2}). \end{aligned} \tag{55}$$

Next, it follows from Theorem 3.18 that, for all $t \geq \hat{t}$,

$$\mathbb{P}[T_\epsilon > t] \leq \exp\left(-\frac{(p-\hat{p})^2 t}{2p^2}\right) + \exp\left(-\min\left\{\frac{s^2 t}{2(2\tau_{-1}\nu)^2}, \frac{st}{2(2\tau_{-1}b)}\right\}\right). \tag{56}$$

Using the tail sum formula for expectation, one has that $\mathbb{E}[T_\epsilon] = \sum_{k=0}^{\infty} \mathbb{P}[T_\epsilon > k]$, which one can split at the ceiling of $\hat{t}$ to obtain

$$\mathbb{E}[T_\epsilon] = \sum_{k=0}^{\lceil \hat{t} \rceil} \mathbb{P}[T_\epsilon > k] + \sum_{m=1}^{\infty} \mathbb{P}[T_\epsilon > \lceil \hat{t} \rceil + m] \leq \lceil \hat{t} \rceil + 1 + \sum_{m=1}^{\infty} \mathbb{P}[T_\epsilon > \lceil \hat{t} \rceil + m]. \tag{57}$$

Next, let us define $c_1 := \frac{(p-\hat{p})^2}{2p^2} > 0$ and $c_2 := \min\left\{\frac{s^2}{2(2\tau_{-1}\nu)^2}, \frac{s}{2(2\tau_{-1}b)}\right\} > 0$, which are motivated by (56). One can use (54) and (51) to obtain

$$c_1 = \frac{(p-\hat{p})^2}{2p^2} = \frac{(\sigma/3)^2}{2p^2} = \frac{\sigma^2}{18p^2} > \frac{(1-\theta)^2(p-\frac{1}{2})^2}{18p^2} =: \underline{c_1}. \tag{58}$$

To bound c_2, we must compute appropriate bounds for s, ν, and b. To bound s, it follows from (52), (48), and (51) that

$$\begin{aligned} s &= \frac{\sigma h(\bar{\alpha})}{3} = \frac{\sigma}{3}\frac{\omega(\bar{\epsilon}_f, 2\bar{\epsilon}_f)}{\rho} \\ &\geq \frac{(1-\theta)(p-\frac{1}{2})}{3}\frac{(2+2\tau_{-1})\bar{\epsilon}_f}{\theta(p-\frac{1}{2})} = \frac{(1-\theta)}{3}\frac{(2+2\tau_{-1})\bar{\epsilon}_f}{\theta}. \end{aligned} \tag{59}$$

Next, since $p > 1/2$, we know that $1/2 + p > 1 - \delta$, which implies that $1/2 > 1 - \delta - p$, which in turn implies that $2 < 1/(1-\delta-p)$, which in turn implies that

$$L := \ln(1/(1-\delta-p)) > \ln(2). \tag{60}$$

We can now also obtain from how p is defined in (17) that

$$L = \ln\left(\frac{1}{1-\delta-p}\right) \equiv \min\left\{\frac{u_f^2}{2\nu^2}, \frac{u_f}{2b}\right\}. \tag{61}$$

It follows from (61) and the definition of u_f in Assumption 3.3 that

$$L \leq \frac{u_f^2}{2\nu^2} < \frac{\bar{\epsilon}_f^2}{2\nu^2} \quad \text{and} \quad L \leq \frac{u_f}{2b} < \frac{\bar{\epsilon}_f}{2b},$$

which after rearranging and using (60) gives

$$\nu < \frac{\bar{\epsilon}_f}{\sqrt{2L}} < \frac{\bar{\epsilon}_f}{\sqrt{2\ln(2)}} = \frac{\bar{\epsilon}_f}{\sqrt{\ln(4)}} \quad \text{and} \quad b < \frac{\bar{\epsilon}_f}{2L} < \frac{\bar{\epsilon}_f}{2\ln(2)} = \frac{\bar{\epsilon}_f}{\ln(4)}.$$

Using these bounds on ν and b, and the bound on s in (59), one has

$$\frac{s}{\nu} \geq \frac{\sqrt{\ln(4)}(1-\theta)(2+2\tau_{-1})}{3\theta} \quad \text{and} \quad \frac{s}{b} \geq \frac{\ln(4)(1-\theta)(2+2\tau_{-1})}{3\theta}.$$

These bounds may now be combined with the definition of c_2 to obtain

$$\begin{aligned} c_2 &= \min\left\{\frac{s^2}{2(2\tau_{-1}\nu)^2}, \frac{s}{2(2\tau_{-1}b)}\right\} \\ &\geq \min\left\{\frac{\ln(4)(1-\theta)^2(1+\tau_{-1})^2}{18\tau_{-1}^2\theta^2}, \frac{\ln(4)(1-\theta)(1+\tau_{-1})}{6\tau_{-1}\theta}\right\} =: \underline{c_2}. \end{aligned} \tag{62}$$

Since the exponents in the bound (56) are linear in t, one has

$$\begin{aligned}
&\sum_{m=1}^{\infty} \mathbb{P}[T_\epsilon > \lceil \hat{t} \rceil + m] \\
\leq &\sum_{m=1}^{\infty} \exp(-c_1(\lceil \hat{t} \rceil + m)) + \sum_{m=1}^{\infty} \exp(-c_2(\lceil \hat{t} \rceil + m)) \\
\leq &\sum_{m=1}^{\infty} \exp(-c_1 m) + \sum_{m=1}^{\infty} \exp(-c_2 m) = \frac{\exp(-c_1)}{1-\exp(-c_1)} + \frac{\exp(-c_2)}{1-\exp(-c_2)} \\
= &\frac{1}{\exp(c_1)-1} + \frac{1}{\exp(c_2)-1} \leq \frac{1}{\exp(\underline{c_1})-1} + \frac{1}{\exp(\underline{c_2})-1},
\end{aligned}$$

where one has that this upper bound is independent of ϵ since both $\underline{c_1}$ and $\underline{c_2}$ are independent of ϵ. One can now combine this result with (57) and (55) to obtain

$$\mathbb{E}[T_\epsilon] \leq \lceil \hat{t} \rceil + 1 + \frac{1}{\exp(\underline{c_1})-1} + \frac{1}{\exp(\underline{c_2})-1} = O(\epsilon^{-2}).$$

The desired result follows by combining this inequality with Lemma 3.4. □ □

4 Conclusion

In this paper, we analyzed a proximal-gradient method for solving regularized stochastic optimization problems with general constraints, where exact objective function and gradient evaluations are inaccessible. We established a high-probability complexity bound under assumptions that accommodate noisy tail distributions commonly encountered in machine learning applications. Our analysis yields three main results: (i) explicit high-probability iteration complexity bounds for both subexponential and bounded-noise settings (Theorem 3.18 and the discussion following that theorem), (ii) almost-sure termination guarantee that an ϵ-KKT point is found in a finite number of iterations (Corollary 3.19(i)), and (iii) an expected iteration complexity of $O(\epsilon^{-2})$ (Corollary 3.19(ii)). These results extend the high-probability analysis framework found in [14] from unconstrained smooth problems to the setting of constrained nonsmooth problems.

5 Probabilistic tools

This section contains some probabilistic results needed to handle the stochastic gradient estimates. We first provide standard closure properties of one-sided subexponential random variables under affine transformations and summation.

Lemma 5.1. *Let X and Y be one-sided subexponential random variables with strictly positive parameters (ν_1, b_1) and (ν_2, b_2), respectively; i.e.,*

$$\begin{aligned}
\mathbb{E}\big[e^{\lambda(X-\mathbb{E}[X])}\big] &\leq e^{\lambda^2\nu_1^2/2} \quad \textit{for all } \lambda \in [0, 1/b_1] \textit{ and} \\
\mathbb{E}\big[e^{\lambda(Y-\mathbb{E}[Y])}\big] &\leq e^{\lambda^2\nu_2^2/2} \quad \textit{for all } \lambda \in [0, 1/b_2].
\end{aligned}$$

Then, the following results hold:

(i) For any $a > 0$ and $c \in \mathbb{R}$, $aX + c$ is one-sided subexponential with parameters $(a\nu_1,\, ab_1)$.

(ii) $X + Y$ is one-sided subexponential with parameters $\big(\sqrt{2(\nu_1^2+\nu_2^2)},\, \max\{2b_1, 2b_2\}\big)$.

Proof. For part (i), we bound the moment-generating function of $aX + c$ as

$$\mathbb{E}\big[e^{\lambda(aX+c-\mathbb{E}[aX+c])}\big] = \mathbb{E}\big[e^{a\lambda(X-\mathbb{E}[X])}\big] \leq e^{a^2\lambda^2\nu_1^2/2} \quad \text{for all } \lambda \in \Big[0, \frac{1}{ab_1}\Big],$$

where the above inequality uses the fact that X is one-sided subexponential with parameters (ν_1, b_1). This proves part (i).

To prove part (ii), let $\lambda \in \Big[0, \min\Big\{\frac{1}{2b_1}, \frac{1}{2b_2}\Big\}\Big] = [0, \frac{1}{\max\{2b_1, 2b_2\}}]$ so that $2\lambda \in [0, 1/b_1] \cap [0, 1/b_2]$. The Cauchy-Schwarz inequality for expectations gives that

$$\begin{aligned}
&\mathbb{E}\big[e^{\lambda(X+Y-\mathbb{E}[X+Y])}\big] \\
&\leq \Big(\mathbb{E}\big[e^{2\lambda(X-\mathbb{E}[X])}\big]\,\mathbb{E}\big[e^{2\lambda(Y-\mathbb{E}[Y])}\big]\Big)^{1/2} \\
&\leq \Big(e^{4\lambda^2\nu_1^2/2}\, e^{4\lambda^2\nu_2^2/2}\Big)^{1/2} = e^{\lambda^2(\nu_1^2+\nu_2^2)} \text{ for all } \lambda \in \Big[0, \tfrac{1}{\max\{2b_1, 2b_2\}}\Big],
\end{aligned}$$

which completes the proof of part (ii). □ □

The next result is a one-sided Bernstein's inequality.

Lemma 5.2. *If X is a random variable satisfying, with strictly positive parameters (ν, b),*

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{\frac{\lambda^2\nu^2}{2}} \quad \textit{for all } \lambda \in \left[0, \frac{1}{b}\right], \tag{63}$$

then, for any $t \geq 0$, it holds that

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \exp\left(-\min\left\{\frac{t^2}{2\nu^2}, \frac{t}{2b}\right\}\right). \tag{64}$$

Proof. If $t = 0$, the right-hand side of (64) is equal to 1 so that the bound holds trivially. Therefore, for the remainder of the proof we suppose that $t > 0$. By combining Markov's inequality and (63) we find, for all $\lambda \in [0, 1/b]$, that

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \mathbb{P}\left(e^{\lambda(X-\mathbb{E}[X])} \geq e^{\lambda t}\right) \leq e^{-\lambda t}\,\mathbb{E}\Big[e^{\lambda(X-\mathbb{E}[X])}\Big] \leq \exp\Big(\frac{\lambda^2\nu^2}{2} - \lambda t\Big).$$

To obtain the tightest bound, we minimize $h_b(\lambda) = \frac{\lambda^2\nu^2}{2} - \lambda t$ over $\lambda \in [0, 1/b]$. Since $\nu^2 > 0$, h_b is a strongly convex quadratic function with *unconstrained* minimizer at $\lambda_u = t/\nu^2 > 0$, which follows from solving $h_b'(\lambda) = \lambda\nu^2 - t = 0$ for λ. We now consider two cases depending on whether λ_u satisfies the constraint $\lambda \leq 1/b$. If $t \leq \nu^2/b$, then $\lambda^* = \lambda_u \in (0, 1/b]$ so that $h_b(\lambda^*) = -t^2/(2\nu^2)$. On the other hand, if $t > \nu^2/b$, then $\lambda^* = 1/b$. Since $\nu^2/b < t$ implies $\nu^2/(2b^2) < t/(2b)$, it follows that $h_b(\lambda^*) = \frac{\nu^2}{2b^2} - \frac{t}{b} \leq \frac{t}{2b} - \frac{t}{b} = -\frac{t}{2b}$. Taking both of these cases into account, we find for all $\lambda \in [0, 1/b]$ that

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \exp\Big(\max\{-\tfrac{t^2}{2\nu^2}, -\tfrac{t}{2b}\}\Big) = \exp\Big(-\min\{\tfrac{t^2}{2\nu^2}, \tfrac{t}{2b}\}\Big),$$

which completes the proof. □ □

References

[1] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

[2] Albert S Berahas, Miaolan Xie, and Baoyu Zhou. A sequential quadratic programming method with high-probability complexity bounds for nonlinear equality-constrained stochastic optimization. *SIAM Journal on Optimization*, 35(1):240–269, 2025.

[3] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Robust sample average approximation. *Mathematical Programming*, 171(1):217–282, 2018.

[4] Digvijay Boob, Qi Deng, and Guanghui Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, 197(1):215–279, 2023.

[5] Digvijay Boob, Qi Deng, and Guanghui Lan. Level constrained first order methods for function constrained optimization. *Mathematical Programming*, 209(1):1–61, 2025.

[6] Changan Chen, Frederick Tung, Naveen Vedula, and Greg Mori. Constraint-aware deep neural network compression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 400–415, 2018.

[7] Frank E Curtis, Xiaoyi Qu, and Daniel P Robinson. A proximal-gradient method for solving regularized optimization problems with general constraints. *arXiv preprint arXiv:2512.23166*, 2025.

[8] Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34:4883–4895, 2021.

[9] Yutong Dai, Xiaoyi Qu, and Daniel P Robinson. A proximal-gradient method for equality constrained optimization. *SIAM Journal on Optimization*, 35(4):2654–2683, 2025.

[10] Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. High-probability complexity bounds for non-smooth stochastic convex optimization with heavy-tailed noise. *Journal of Optimization Theory and Applications*, 203(3):2679–2738, 2024.

[11] Mert Gürbüzbalaban, Asuman Ozdaglar, and Pablo A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186:49–84, 2021.

[12] Basil M Idrees, Lavish Arora, and Ketan Rajawat. Constrained stochastic recursive momentum successive convex approximation. *IEEE Transactions on Signal Processing*, 2025.

[13] Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. In *Conference on Learning Theory*, pages 1752–1755. PMLR, 2019.

[14] Billy Jin, Katya Scheinberg, and Miaolan Xie. High probability complexity bounds for adaptive step search based on stochastic oracles. *SIAM Journal on Optimization*, 34(3):2411–2439, 2024.

[15] Lingzi Jin and Xiao Wang. A stochastic primal-dual method for a class of nonconvex constrained optimization. *Computational Optimization and Applications*, 83(1):143–180, 2022.

[16] Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive SGD with momentum. In *Workshop on Beyond First Order Methods in ML Systems at the 37th International Conference on Machine Learning (ICML)*, 2020.

[17] Langqi Liu, Yibo Wang, and Lijun Zhang. High-probability bound for non-smooth non-convex stochastic optimization with heavy tails. In *Forty-first International Conference on Machine Learning*, 2024.

[18] Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[19] Lam M. Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takáč. SGD and Hogwild! convergence without the bounded gradients assumption. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3750–3758. PMLR, 2018.

[20] R. Tyrrell Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1970.

[21] Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, pages 29563–29648. PMLR, 2023.

[22] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

[23] Qiankun Shi, Xiao Wang, and Hao Wang. A momentum-based linearized augmented lagrangian method for nonconvex constrained stochastic optimization. *Mathematics of Operations Research*, 2025.

[24] Vikrant Singh Tomar and Richard C Rose. Manifold regularized deep neural networks. In *INTERSPEECH*, pages 348–352, 2014.

[25] Yinhao Zhu, Nicholas Zabaras, Phaedon-Stelios Koutsourelakis, and Paris Perdikaris. Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *Journal of computational physics*, 394:56–81, 2019.